UNIVERSITY OF THE AEGEAN

Department of Statistics and
Actuarial-Financial Mathematics

Statistics and Data Analysis

Region Charts for Time Series Monitoring

MASTER THESIS

KONSTANTINOS

LADOPOULOS

SAMOS 2019

EVALUATION COMMITTEE

Karagrigoriou Alex
Professor UAegean (Supervisor)

Rakitzis Athanasios
Assistant Professor UAegean

Xatzopoulos Petros
Assistant Professor UAegean

Contents

Abstract

This thesis is conducted at the *Department of Statistics and Actuarial-Financial Mathematics* of the *University of the Aegean*, in the context of the MSc program in *Statistics and Data Analysis*. Its purpose, is to analyze and evaluate the most commonly used transformation methods in off-line quality control. To that end, among others, Box-Cox and Yeo-Johnson methods are presented along with performance measures and fundamental elements of statistical quality control. Based on the aforementioned, a new transformation approach for time-series monitoring data is proposed.

The structure of the thesis consists of four Chapters. Chapter 1 constitutes an introduction to statistical quality control. Thus, the basics of the latter are presented. Control charts such as $\overline{X} - S$ and $\overline{X} - R$ for monitoring process statistics as well as joint schemes for monitoring two statistics simultaneously, are also discussed.

In Chapter 2, the need of power transformations in statistical quality control is raised. Therefore, various power transformations (i.e., Box-Cox, Yeo-Johnson, Logothetis, etc.) as well as noise and target performance measures are presented.

In Chapter 3, Analysis of Variance (ANOVA) is described and properly implemented in Experimental Designs. Among others, Two-Way ANOVA, Two-Way Repeated Measures ANOVA and mixed models are presented.

Finally, in Chapter 4, new adjusted transformation methods for off-line quality control are proposed, which result in proper performance measures for the determination of the controllable factors that affect the mean and variability of the response variable of interest. The proposed adjusted transformations are compared with the well-known transformations of Box-Cox and Logothetis. The performance abilities of the proposed methodology are demonstrated on quality control data, considering both a real dataset and a simulated one.

Περίληψη

Η παρούσα διατριβή εκπονήθηκε στο *Τμήμα Στατιστικής και Αναλογιστικών - Χρηματοικονομικών Μαθηματικών* του *Πανεπιστημίου Αιγαίου*, στα πλαίσια του Προγράμματος Μεταπτυχιακών Σπουδών "*Στατιστική και Ανάλυση Δεδομένων*". Σκοπός της είναι η ανάλυση και αξιολόγηση μεθόδων μετασχηματισμού που χρησιμοποιούνται συχνά για τον έλεγχο της ποιότητας ενός προϊόντος πριν την παραγωγή (off-line Quality Control). Για τον σκοπό αυτό, μεταξύ άλλων, παρουσιάζονται οι μετασχηματισμοί των Box-Cox και Yeo-Johnson μαζί με μέτρα επίδοσης καθώς και θεμελιώδη στοιχεία του στατιστικού ελέγχου ποιότητας. Με βάση τα προαναφερθέντα, προτείνεται μία νέα προσέγγιση μετασχηματισμού για την παρακολούθηση χρονολογικών δεδομένων.

Η δομή της διατριβής αποτελείται από τέσσερα Κεφάλαια. Το Κεφάλαιο 1 αποτελεί εισαγωγή στον στατιστικό έλεγχο ποιότητας. Έτσι, τα βασικά του χαρακτηριστικά παρουσιάζονται. Επιπλέον, παρουσιάζονται Διαγράμματα ελέγχου όπως τα $\overline{X} - S$ και $\overline{X} - R$, για την παρακολούθηση στατιστικών διεργασίας καθώς και από κοινού διαγράμματα για την ταυτόχρονη παρακολούθηση δύο στατιστικών.

Στο Κεφάλαιο 2, περιγράφεται η χρησιμότητα του μετασχηματισμού δεδομένων πάνω στο στατιστικό έλεγχο ποιότητας. Επομένως, παρουσιάζονται διάφοροι μετασχηματισμοί (όπως αυτοί των Box-Cox, Yeo-Johnson, Logothetis κλπ.), καθώς και μέτρα επίδοσης θορύβου και στόχου.

Στο Κεφάλαιο 3, γίνεται περιγραφή και εφαρμογή της Ανάλυσης Διακύμανσης (ANOVA) σε Πειραματικούς Σχεδιασμούς. Μεταξύ άλλων, παρουσιάζονται πειραματικοί σχεδιαμοί για Ανάλυση Διακύμανσης Δύο Παραγόντων (Two-Way ANOVA), Ανάλυση Διακύμανσης Δύο Παραγόντων με Επαναλαμβανόμενες Μετρήσεις (Two-Way Repeated Measures ANOVA), καθώς και μεικτά μοντέλα (Mixed Models).

Τέλος στο Κεφάλαιο 4, παρουσιάζονται νέες προσαρμοσμένες μέθοδοι μετασχηματισμού δεδομένων για τον έλεγχο ποιότητας πριν την παραγωγή, οι οποίες με την σειρά τους οδηγούν σε κατάλληλα μέτρα απόδοσης για τον προσδιορισμό ελεγχόμενων παραγόντων οι οποίοι επηρεάζουν τον μέσο όρο και την μεταβλητότητα όσον αφορά της υπό μελέτη μεταβλητής απόκρισης. Στην συνέχεια πραγματοποιείται σύγκριση μεταξύ των προτεινόμενων μεθόδων, με τους μετασχηματισμούς Box-Cox και Logothetis. Η ικανότητα απόδοσης των προτεινόμενων μεθοδολογιών παρουσιάζεται πάνω σε πραγματικά και προσομοιωμένα δεδομένα ελέγχου ποιότητας.

Chapter 1

Control Charts

With the rise of industrial revolution, the majority of companies aim in constantly improving the quality of their product in order to be reliable and competitive in the market. In this Chapter, basic elements of statistical quality control will be discussed. More specifically, three classical statistical quality control charts will be presented which are frequently used by decision makers of a company for monitoring charts, the behaviour of the products line.

1.1 Introduction

Nowadays, human consuming behaviours depend on the price and the quality of a product or service. In order to achieve so, a company or business must adopt various techniques for quality control.

Statistical quality control (SQC) is one of the most important techniques used for decision making regarding the production and quality of products. With the use of SQC, defective products can be detected and hence, proper actions could take over and remove them in order to maintain the quality of the product line.

Most companies or businesses aim in constantly improving the quality of their product in order to produce reduced or ideally zero errors. To achieve so, they monitor products, via properly trained staff, for significant variations in order to increase the quality control.

Quality has a multidimensional meaning. A simple definition of quality is given by Joseph Juran in [18] "quality means the fitness for use". More specifically, if a product satisfies the needs of a marketplace, it is referred as quality product. In statistical terms "quality is mainly determined by the amount of variability in what is being measured". Depending on consumer preferences, some of the product quality dimensions are: reliability, output, aesthetics, capability of a product and company reputation. To operate effectively the SQC process, it is considered beneficial for companies to affiliate a plan in order to ensure the constant quality improvement in all sections of a business. The aforementioned plan is known as **Total Quality Management**.

Statistical quality control can be divided into three categories of statistical methods for data analysis:

- Design of Experiments (DoE)
  It is related with detecting which factor levels affect the quality characteristics of a product.

- Statistical Process Control (SPC)
  It is responsible for controlling the production process through statistical techniques.

- Acceptance Sampling (AS)
  Helps in deciding if a sample characteristic that is under examination should be excluded or not from the process.

In a SQC process there exist two main reasons that can cause variation. The first one is *"common variability"* and the causes that lead to this kind of variation are called *common or chance causes of variation*. Regardless of how fine is the raw material of a product or how good are the machine operators, variation will always exist because of uncontrollable factors. When the aforementioned reasons appear, the process is in a *stable state* or *in control* (IC).

The second one is called *"special variability"* and the causes that lead to this kind of variation are called *special or assignable causes of variation* (e.g., incorrectly adjusted machines, poor quality of raw material, etc.). When special variability occurs, then the process is considered *out of control* (OOC) or operates in an *unstable state*.

During the design of a product, the *specifications limits* and the *Target* value **(T)** are determined for the *quality characteristics*. There are two specifications limits i.e., the **lower** and **upper specification limit [LSL, USL]** and between them should fall all quality characteristics so that a product is quality accepted. The value T represents the value that is desirable for maximizing the quality of a product and it is usually located in the center of the interval [LSL, USL].

The products with at least one quality feature outside the specifications limits are called **non conforming products**. There are cases where the size and seriousness of the defects is not big enough and thus, the companies instead of disposing them, let them in market.

In SQC, it is essential to examine the performance ability of a process. To achieve so, a set of statistical techniques is applied known as process capability analysis. By comparing the sample values of a distribution with the specification limits, process capability analysis achieves in identifying the number of non conforming products produced. Procedures like the one above, assumes normality and stability and thus, the quality characteristics must be derived from an IC process.

The use of *control charts* allows the real time monitoring of special causes of variation. Hence, SQC charts are used such as Shewhart, CUSUM and EWMA. Finally, control charts are divided into two main categories based on the quality characteristic under examination:

- Control Charts for attributes described by discrete random variables;

- Control Charts for attributes described by continuous random variables.

For more details on SQC, the interested reader may refer to [2];[10];[25]; and [31].

1.2 Quality Control Chart

1.2.1 Definition

A quality control chart is a special type of graph which reenacts whether the characteristics of the data under examination are meeting the indented specifications or not. It displays values of statistical functions, e.g., mean, variance, standard deviation, etc., as points in a x-y axis system, where x-axis represents the number of samples collected or time and y-axis represents the values of the statistic function. When a control chart analyzes a specific characteristic of a product, it is referred as *"univariate quality control chart"*, while in case of analyzing more than one product characteristic it is referred as "multivariate quality control chart".

A typical quality control chart consists of three horizontal lines: (1) the center line (CL); (2) the lower control limit (LCL); and (3) the upper control limit (UCL), where the CL stands for the mean or average value of the characteristic which is represented in the control chart. If the sample values fall between the control limits, then the process is considered IC, while if at least one point is out of the control limits then the process is considered OOC. In the case of an OOC process, the problem needs to be investigated and finally fixed before a big amount of faulty products are produced.

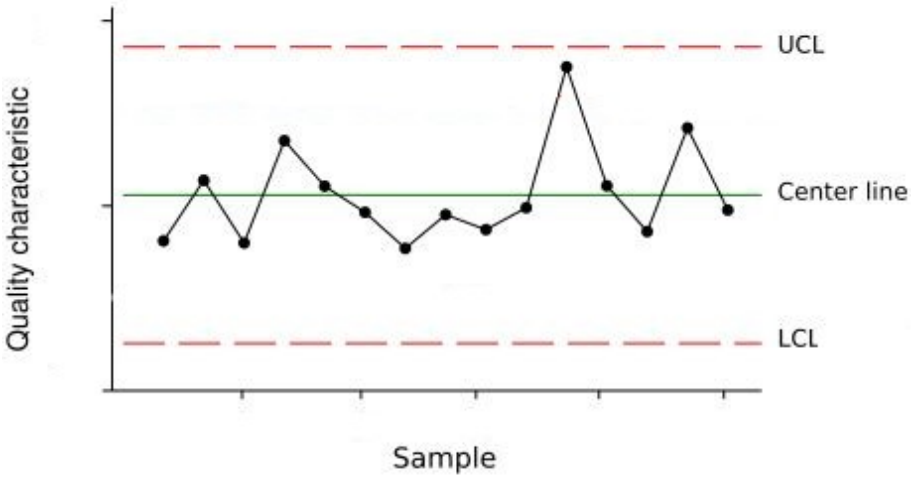

Figure 1.1: A typical quality control chart

In Figure 1.1 the points are randomly distributed and between the LCL and UCL which indicates that the process is IC.

There are cases where the points fall between the two limits but the process is OOC. This occurs when the points are non-randomly distributed and they act in a systematic way. Due to this fact many analysts are using a set of limits known as **warning limits**. This new set of limits are usually designated for two sigma (*outer*) and one-sigma (*inner*) warning limits and indicate if an IC process has to be reexamined. In case of two sigma warning limits, the sensitivity of the control chart is increased and as a result it is easier to identify the cause of possible systematic patterns.

Figure 1.2 represents a $\overline{X}$ control chart (which monitors whether the sample means are IC) with one, two and three sigma control limits creating three zones (A,

B and C).

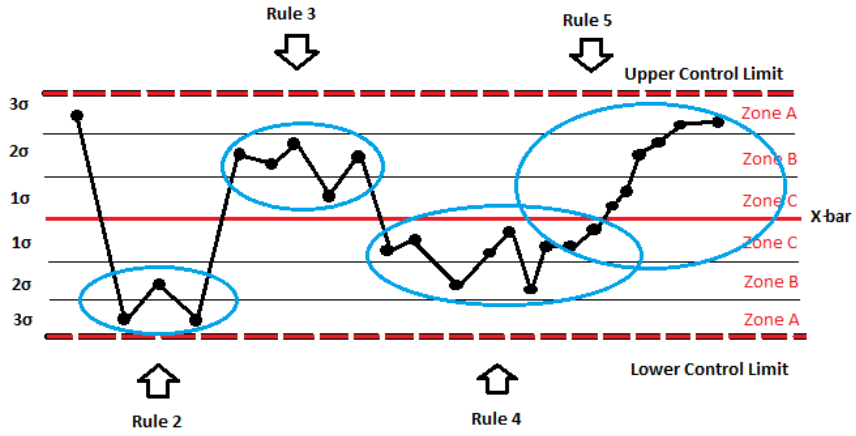

Figure 1.2: Zone rules applied in a $\overline{X}$ Control chart

There are rules where if at least one of them is applied then the process might be OOC:

1. One or more points outside of the control limits.

2. Two out of three consecutive points outside the two-sigma warning limits but still inside the control limits.

3. Four out of five consecutive points beyond the one-sigma limits.

4. A run of eight consecutive points on one side of the center line.

5. Six or more points in a row steadily increasing or decreasing.

6. Fifteen points in a row in zone C (both above and below the center line).

7. Fourteen points in a row alternating up and down.

8. Eight points in a row on both sides of the center line with none in zone C.

9. An unusual or non-random pattern in the data.

10. One or more points near a warning or control limit.

Rules (1-4) are called **Western Electrics** or **Zone Rules** for the control charts.

When the process is OOC, a sequence of activities and decisions must be made in order to maintain the quality of the products and thus, an **out-of-control-plan (OCAP)** must be applied.

1.3 Phase I and Phase II

To properly use a control chart, the researcher/operator must know in which Phase he is currently in, **Phase I** or **Phase II**.

The goal in **Phase I** is to determine and examine if a set of data collected over time is IC. If the process is IC, the constructed limits can be used for future monitoring of the process. If the process is OOC, then the operator first brings it in control and then uses the new pair of limits for future monitoring, which is the purpose that Phase I serves. When points are spotted out of the control limits, they are excluded from the analysis. The most common number of samples that are used in such a procedure is m=20 or m=25. That specific use of the control charts in Phase I is called *retrospective*.

In **Phase II** the control charts are used to constantly monitor if the procedure is in or out of control. New data are collected at regular intervals and using the in control limits of Phase I, the operator decides if the process falls between the two control limits. Different mathematical formulas are used for each phase and that is why the operator must know a-priory in which phase the process must be planned.

1.4 Three-Sigma Control Limits

The Three-Sigma (3σ) rule refers to the data that are within three standard deviations from the mean. Sigma (σ), i.e., the standard deviation, measures the amount of variation that exists between the observed data and the average.
For example the basic equations of control limits are the following:

$$\begin{aligned} LCL &= \mu - L\sigma, \\ UCL &= \mu + L\sigma, \end{aligned} \tag{1.1}$$

where μ, σ are the mean and standard deviation respectively and L represents the distance between the control limits from the center line. In the special case of $L = 3$, the limits presented in equation (1.1) are called **Three-Sigma Control Limits (TSCL)**. Approximately for a normally distributed dataset the 99.7% of the values lie underneath the curve and it can be expressed as follows:

$$Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973, \tag{1.2}$$

that is the probability of a value exceeding the TSCL (*"false alarm"*) is 0.27%.

1.5 $\overline{X} - R$ Control Chart

The $\overline{X} - R$ charts constitute a pair of charts that are used to plot and monitor the behavior of the mean ($\overline{X}$) and the range (R) over time for continuous data. The $\overline{X}$ control limits are calculated via the $R - chart$ and as a result if a $R-$chart is OOC then the control limits of the $\overline{X}-$chart are incorrect and display Type I or Type II error. Typically $\overline{X} - R$ charts are used when the sample or subgroup size is between two and ten.

1.5.1 Statistical Formulas

In order to construct control charts, the population parameters μ and σ need to be estimated from the precursory samples. Let us suppose that there are m samples (usually twenty to twenty-five in number) and each sample has n observations (usually the number of observations varies from four to six). Let also $\overline{x}_1, \overline{x}_2, ..., \overline{x}_m$ be the average value of each sample. Then the overall average $\overline{\overline{x}}$ is the best estimator of μ and is given by the following equation:

$$\overline{\overline{x}} = \frac{\overline{x}_1 + \overline{x}_2 + ... + \overline{x}_n}{m}. \tag{1.3}$$

Note that $\overline{\overline{x}}$ constitutes the CL of the $\overline{X}-$chart.

To construct the control limits of the $\overline{X} - R$ chart an estimation of the standard deviation σ must be available. The estimation can be derived through the standard deviations or ranges of the m samples. In our case, since we try to construct the $\overline{X} - R$ chart, we will make use of the range method. Let $x_1, x_2, ..., x_n$ denote the observations within a sample. Then the range is defined as:

$$R = x_{max} - x_{min}. \tag{1.4}$$

For the estimation of the control limits we need the average value of all ranges among the m samples. To do so we make use of the expression

$$\overline{R} = \frac{R_1 + R_2 + ... + R_m}{m}. \tag{1.5}$$

1.5.2 R-Chart Control Limits

The TSCL for the $R - chart$ (Phase I) are given by:

$$\overline{R} \pm 3\widehat{\sigma}_R. \tag{1.6}$$

If the characteristic comes from a normal distribution, $\widehat{\sigma}_R$ can be calculated from the relative range defined as:

$$W = \frac{R}{\sigma'}. \tag{1.7}$$

The estimator of the standard deviation for R is given by:

$$\widehat{\sigma}_R = d_3\sigma', \tag{1.8}$$

where d_3 represents the estimator for the standard deviation of W and $\sigma' = \overline{R}/d_2$. Hence, equation (1.8) can be rewritten as :

$$\widehat{\sigma}_R = d_3\frac{\overline{R}}{d_2}, \tag{1.9}$$

where d_2 is an adjustment factor used for the estimation of the standard deviation and its values depend on the sample that is used. Note that d_2 and d_3 can be calculated via Duncan Table M, Appendix, p.886 [10].

In conclusion, the TSCL and the CL for the $R-$chart can be calculated from the equations bellow:

$$\begin{aligned} LCL &= \overline{R} - 3\widehat{\sigma}_R = \overline{R} - 3d_3\frac{\overline{R}}{d_2} = D_3\overline{R}, \\ CL &= \overline{R}, \\ UCL &= \overline{R} + 3\widehat{\sigma}_R = \overline{R} + 3d_3\frac{\overline{R}}{d_2} = D_4\overline{R}, \end{aligned} \tag{1.10}$$

where $D_3 = (1-3)d_3/d_2$, $D_4 = (1+3)d_3/d_2$.
For Phase II the TSCL for the $R-$chart are given by the equations

$$\begin{aligned} LCL &= \mu_{R_i} - 3\sigma_{R_i} = (d_2 - 3d_3)\sigma = D_1\sigma, \\ CL &= \mu_{R_i}, \\ UCL &= \mu_{R_i} + 3\sigma_{R_i} = (d_2 + 3d_3)\sigma = D_2\sigma. \end{aligned} \tag{1.11}$$

In the special case of $n \leq 6$ then $D_1 < 0$ and $LCL = 0$.

1.5.3 $\overline{X}$-Chart Control Limits

The TSCL for the $\overline{X}-$chart are given by the following expression:

$$\overline{\overline{X}} \pm 3\widehat{\sigma}_{\overline{x}}, \tag{1.12}$$

where

$$\widehat{\sigma}_{\overline{x}} = \frac{\widehat{\sigma}_x}{\sqrt{n}}. \tag{1.13}$$

The statistic $\widehat{\sigma}_x$ is calculated from the subgroup of ranges and has the following form:

$$\widehat{\sigma}_x = \frac{\overline{R}}{d_2}. \tag{1.14}$$

As for the $\overline{X}$ control limits, with the standard deviation considered as the average of ranges, are presented below

$$\begin{aligned} \overline{\overline{X}} \pm 3\widehat{\sigma}_{\overline{x}} &= \overline{\overline{X}} \pm 3\frac{\widehat{\sigma}_x}{\sqrt{n}} \\ &= \overline{\overline{X}} \pm 3\frac{\left(\overline{R}/d_2\right)}{\sqrt{n}} \\ &= \overline{\overline{X}} \pm A_2\overline{R}, \end{aligned} \tag{1.15}$$

where $A_2 = 3/d_2\sqrt{n}$.

Finally the TSCL and CL for the $\overline{X}-$chart are as follows:

$$\begin{aligned} LCL &= \overline{\overline{X}} - A_2\overline{R}, \\ CL &= \overline{\overline{X}}, \\ UCL &= \overline{\overline{X}} + A_2\overline{R}. \end{aligned} \tag{1.16}$$

1.6 $\overline{X} - S$ Control Chart

Although the $R-$chart is widely used for many years, it has a defect: when the subgroup size increases, it cannot explain adequately the within group variation. Thus, for large subgroup sizes, $\overline{X} - S$ control charts are preferable since they give better solutions regarding the variation of larger subgroup sizes. They are used for continuous variables (e.g., time, weight, length, etc.) and are composed by two charts: one chart for the representation of $\overline{X}$ and another one for the representation of S.

1.6.1 S-Chart Control Limits

When the value of σ is known (Phase II), $E(S) = c_4\sigma$ represents the CL and the TSCL can be calculated as follows:

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

where c_4 represents a constant, which depends on the sample size n and $\sigma\sqrt{1 - c_4^2}$ is the standard deviation of S. The above equations can be also written as:

$$\begin{aligned} LCL &= B_5\sigma, \\ CL &= c_4\sigma, \\ UCL &= B_6\sigma, \end{aligned} \tag{1.17}$$

where

$$\begin{aligned} B_5 &= c_4 - 3\sqrt{1 - c_4^2}, \\ B_6 &= c_4 + 3\sqrt{1 - c_4^2}. \end{aligned} \tag{1.18}$$

In addition the control limits and the center line for $\overline{X}$ can be expressed as:

$$\begin{aligned} LCL &= \mu - A\sigma, \\ CL &= \mu, \\ UCL &= \mu + A\sigma, \end{aligned} \tag{1.19}$$

where $A = 3/\sqrt{n}$.

When the values of a characteristic are normally distributed and the σ is unknown (Phase I), the TSCL for the $S-$chart are given by the formula:

$$\overline{S} \pm 3\widehat{\sigma}_s, \tag{1.20}$$

where $\overline{S}$ is the average of all the subgroup standard deviations and it is calculated by

$$\overline{S} = \frac{1}{m}\sum_{i=1}^{m} s_i. \tag{1.21}$$

We previously discussed about the standard deviation of (S) which is equal to

$$\sigma\sqrt{1-c_4^2}, \tag{1.22}$$

where $\widehat{\sigma} = \overline{S}/c_4$. Thus, the TSCL and the CL of the $S-$chart for Phase I takes the form:

$$\begin{aligned} LCL &= \overline{S} - 3\frac{\overline{S}}{c_4}\sqrt{1-c_4^2}, \\ CL &= \overline{S}, \\ UCL &= \overline{S} + 3\frac{\overline{S}}{c_4}\sqrt{1-c_4^2}. \end{aligned}$$

If $B_3 = 1 - \frac{3}{c_4}\sqrt{1-c_4^2}$ and $B_4 = 1 + \frac{3}{c_4}\sqrt{1-c_4^2}$, then the $S-$chart control limits can be written as:

$$\begin{aligned} LCL &= B_3\overline{S}, \\ CL &= \overline{S}, \\ UCL &= B_4\overline{S}. \end{aligned} \tag{1.23}$$

1.6.2 $\overline{X}$-Chart Control Limits

Having the estimator $\overline{S}/c_4$ one can calculate the TSCL and the CL of the $\overline{X}-$chart as follows:

$$\begin{aligned} LCL &= \overline{\overline{X}} - \frac{3\overline{S}}{c_4\sqrt{n}}, \\ CL &= \overline{\overline{X}}, \\ UCL &= \overline{\overline{X}} + \frac{3\overline{S}}{c_4\sqrt{n}}. \end{aligned}$$

In addition, if $A_3 = 3/(c_4\sqrt{n})$ the above equations can be written as:

$$\begin{aligned} LCL &= \overline{\overline{X}} - A_3\overline{S}, \\ CL &= \overline{\overline{X}}, \\ UCL &= \overline{\overline{X}} + A_3\overline{S}. \end{aligned} \tag{1.24}$$

Example (Montgomery, (2009), p.260). The dataset concerns piston rings diameter and contain forty samples ($m = 40$) with five observations each ($n = 5$). We are going to use the first 30 samples for Phase I and the rest 10 samples for Phase II. The TSCL for the $\overline{X}-$S chart in Phase I are calculated via equations (1.23) and (1.24), respectively (see Fig 1.3 and 1.4).

We observe that all the points in Figures 1.3 and 1.4 fall between the control limits. Hence, both processes are in control and we can use their control limits for future monitoring. For Phase II, the same control limits from Phase I will be used and the graphs will be plotted again with the starter values and the rest ten samples.

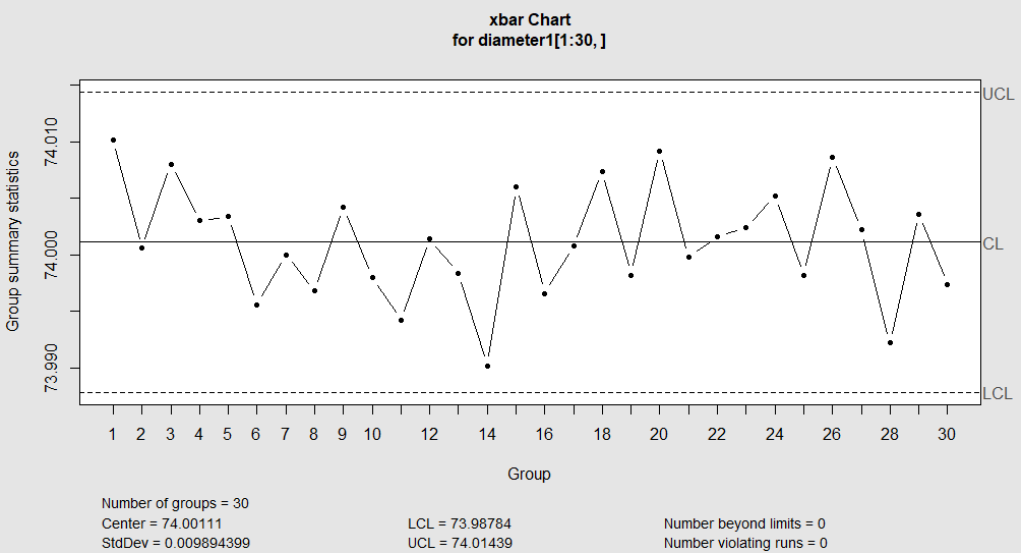

Figure 1.3: $\overline{X}-$chart for pistonrings dataset with TSCL, including the first $m = 30$ samples (Phase I)

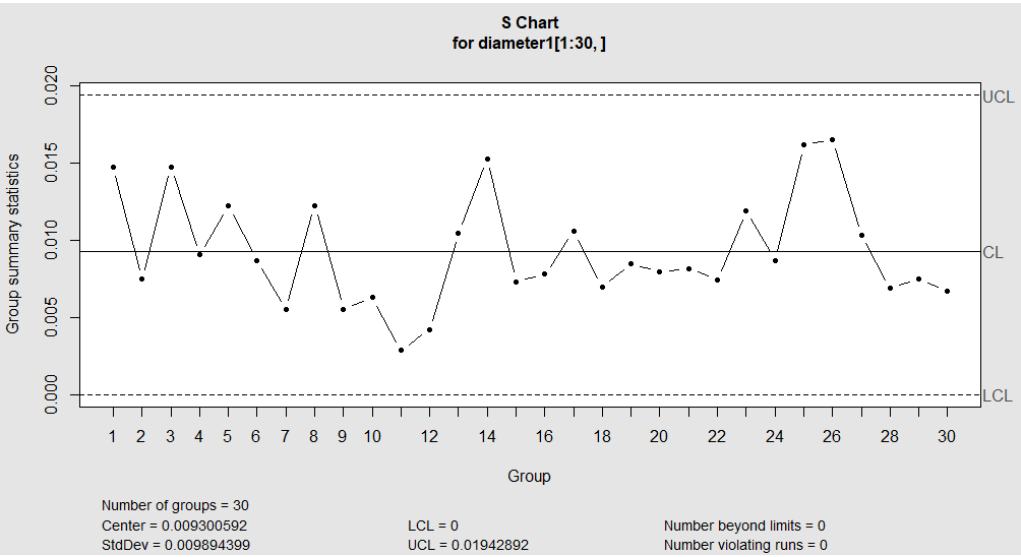

Figure 1.4: $S-$chart for pistonrings dataset with TSCL, including the first $m = 30$ samples (Phase I)

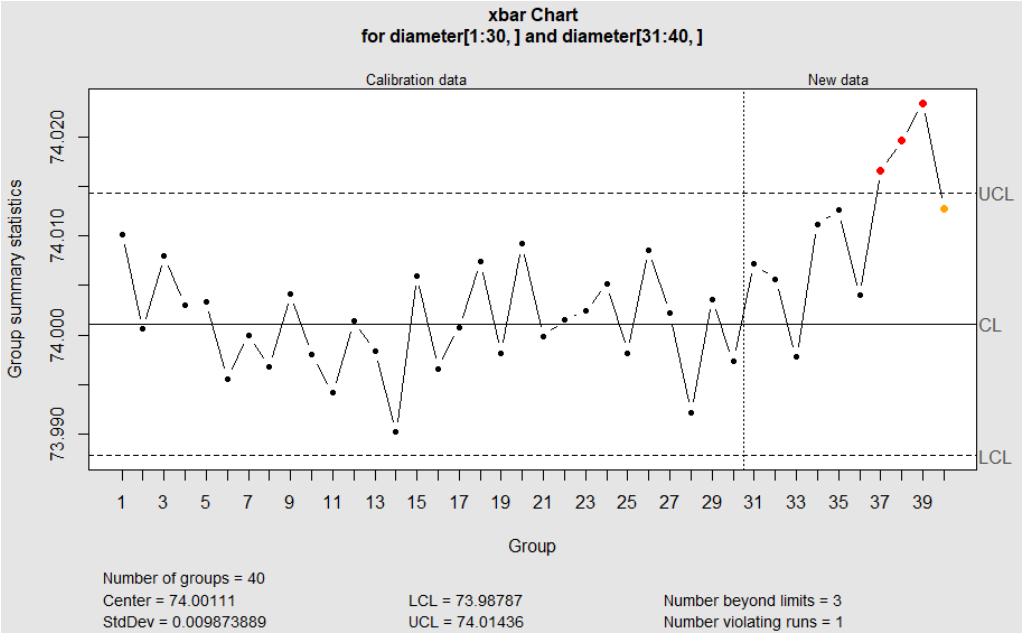

Figure 1.5: $\overline{X} - chart$ for pistonrings dataset with TSCL, including all samples $m = 40$ (Phase II)

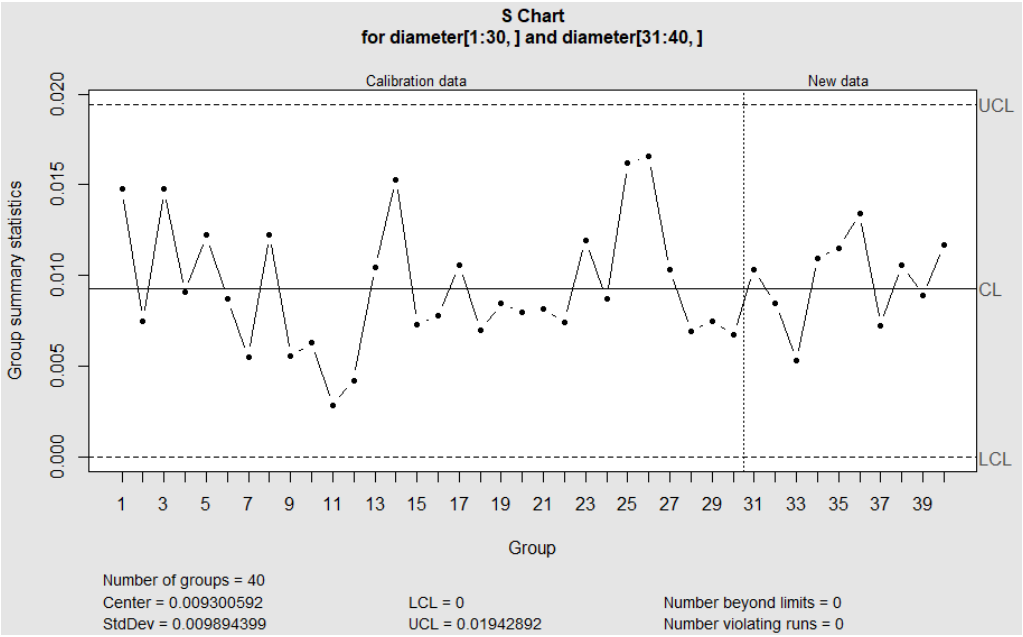

Figure 1.6: $S - chart$ for pistonrings dataset with TSCL, including all samples $m = 40$ (Phase II)

Once again, the process for the $S-$chart for $n = 40$ samples remains in control. However, in the $\overline{X}-$chart two points (in red color) fall outside the control limits. In this case we must find the reason why this shift occurred and fix it.
The points that fall outside the control limits will be omitted and new control limits will be created. The aforementioned will be repeated until the process becomes stable.

1.7 Individuals and Moving-Range $(I-MR)$ Control Charts

When the subgroup size is $n=1$, $I-MR$ control charts are used for monitoring the variation of the process and the individuals values for continuous over time data.

Let us suppose that the characteristic under investigation is normally distributed and a random sample $X_1, X_2, ..., X_m$ is available. For the $I-$chart we can use the MR of two consecutive observations which is defined as:

$$MR_i = |X_i - X_{i-1}|, \ \ 2 \leq i \leq m, \tag{1.25}$$

where

$$\mu_{MR_i} = E(MR_i) = \sigma d_2, \ \ \sigma_{MR_i} = \sigma d_3.$$

For Phase I

$$\widehat{\mu} = \overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \ \ \widehat{\sigma} = \overline{MR}/d_2, \tag{1.26}$$

where

$$\overline{MR} = (MR_2 + MR_3 + ... + MR_m)/(m-1). \tag{1.27}$$

The TSCL for Phase I and the CL for the $MR-$chart are given by

$$\overline{MR} \pm 3\widehat{\sigma}_{MR}, \tag{1.28}$$

where $\widehat{\sigma}_{MR} = \widehat{\sigma} d_3$.

According to equation (1.26), the estimation of $\widehat{\sigma}_{MR}$ can be rewritten as

$$\widehat{\sigma}_{MR} = \frac{d_3}{d_2}\overline{MR}. \tag{1.29}$$

The TSCL for the $MR-$chart are given by the following formulas:

$$\begin{aligned} LCL &= \left(1 - 3\frac{d_3}{d_2}\right)\overline{MR} = D_3\overline{MR}, \\ CL &= \overline{MR}, \\ UCL &= \left(1 + 3\frac{d_3}{d_2}\right)\overline{MR} = D_4\overline{MR}, \end{aligned} \tag{1.30}$$

where D_3, D_4 are calculated for n=2 (see Duncan Table M in Appendix, p.886) [10].

The TSCL for the $I-$chart are given by

$$\widehat{\mu} \pm 3\widehat{\sigma}.$$

By replacing the estimators $\widehat{\mu}$ and $\widehat{\sigma}$ the following equations are derived

$$\begin{aligned} LCL &= \overline{X} - 3\frac{\overline{MR}}{d_2}, \\ CL &= \overline{X}, \\ UCL &= \overline{X} + 3\frac{\overline{MR}}{d_2}. \end{aligned} \tag{1.31}$$

In **Phase II**, where μ and σ are known, the TSCL for the $I-MR$ control charts are as follows:
For the $I-$chart:

$$\begin{aligned} LCL &= \mu - 3\sigma, \\ CL &= \mu, \\ UCL &= \mu + 3\sigma, \end{aligned} \tag{1.32}$$

and for the $MR-$chart:

$$\begin{aligned} LCL &= \mu_{MR_i} - 3\sigma_{MR_i} = (d_2 - d_3)\sigma = D_1\sigma, \\ CL &= d_2\sigma, \\ UCL &= \mu_{MR_i} + 3\sigma_{MR_i} = (d_2 + d_3)\sigma = D_2\sigma, \end{aligned} \tag{1.33}$$

where D_1, D_2 are calculated for n=2.

1.8 Joint Monitoring Schemes for Mean and Variance

In various cases, it is preferable to monitor simultaneously the mean and variance of a distribution in a process because where special causes of variability exist, a small shift in the variance can change the control limits of the $\overline{X}-$chart or they can simultaneously change at the same time point. Thus, a number of schemes combine the mean and variance in one chart in order to locate the shifts that occur in specific time intervals in either mean or variance. When normality cannot be assumed, then distribution free or non parametric charts can be applied.

One chart monitoring schemes are quite popular and attractive because of their simplicity. They often contain one charting statistic which is located in a single graph. The charting statistic is a combination of mean and variance, and control limits can be found via the distribution of the statistic.

When the standard deviation and mean are known and they come from normal distribution, the one chart joint monitor schemes are called "case K" or "standards known".

Chart joint monitoring schemes in case K are divided in two classes:

1. simultaneous control charts with two statistics and

2. single control charts with one statistic, which is the combination of the mean and variance.

Furthermore, single control charts can be divided into those with ordinary control limits and those with a two dimensional control region.

In **simultaneous control charts**, there are two different statistics; one for the mean and one for the variance, and they are plotted on the same graph.

The single charting statistic that **single charts with ordinary control limits** use, is usually a function of the minimal sufficient statistics $\overline{X}$ and S^2. The most popular chart in this category is the $Max-$chart proposed by Chen and Cheng in

[7], which considers the maximum of the absolute values of two normalized statistics. The first normalized statistic refers to mean and the other to variance. The combination of these statistics can be written as follows:

$$M = max(|U_i|, |V_i|), \tag{1.34}$$

where

$$U_i = \frac{\overline{X}_i - \mu}{\sigma/\sqrt{n_i}} \quad \text{and} \quad V_i = \Phi^{-1}\left\{H\left[\frac{(n_i-1)S_i^2}{\sigma^2}; n_i - 1\right]\right\}, \tag{1.35}$$

where H denotes the cumulative distribution function (cdf) of the chi-square distribution with $n_i - 1$ degrees of freedom, n_i refers to the size of the i-th sample and Φ denotes the cdf of the standardized normal distribution.

In **Single charts with control regions**, data are represented in a two dimensional plane and if they are spotted within an appointed control region then the process is considered IC otherwise is considered OOC. The problem with such graphs is that trend, which is related with time, cannot be spotted.

A plethora of charts like the aforementioned have been developed with semi-circular, circular or elliptical control regions and each one has its own advantages. For example, rectangular control regions are eligible for changes in mean and elliptical control regions are eligible for both mean and variance.

Chao and Cheng in [6] proposed a control chart which the points $(\overline{X}_i, S_i^*)$ are plotted on the $(\overline{X}_i, S_i^*)$ plane, where

$$S_i^* = \sqrt{(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2}, \quad \text{for } i = 1, 2, ..., n. \tag{1.36}$$

The statistic $T = (\overline{X} - \mu_0)^2 + S^{*2}$ is used for the creation of the control region. When the sample data come from normal distribution, $(n/\sigma_0^2)T$ has a chi-square distribution with n degrees of freedom and the statistic T determines a circular region. Note that S_i^* must be positive and for that reason only the half of the control region is needed. This kind of control chart is called Semi-Circle chart (SC-chart).

Later, Chao and Cheng in [5] constructed a $SC-$chart s.t. minimum coverage area known as EWMA$-SC$ chart. The control region of this chart is not the area under a semi circle, but the area under a line and the T statistic is disintegrated into mean and variance components,

$$U_i = \left[\frac{n(X_i - \mu_0)^2}{\sigma_0^2} - 1\right] \quad \text{and} \quad V_i = \left[(n-1)\left(\frac{S_i^2}{\sigma_0^2} - 1\right)\right]. \tag{1.37}$$

If the sample points are located in the $U - V$ plane and they are below the line, then the process is IC. For a more detailed overview in joint monitoring schemes please refer in [24].

Chapter 2

Power Transformations

In statistics, a power transformation family includes functions that are used to create monotonic transformations of data. The use of power functions for data transformation, helps in achieving constancy of variance and makes skewed distributed data come closer to normal distribution (i.e., more symmetrical). In addition, simplicity or linearity of the model structure can be achieved. In SQC there is a variety of procedures in which normality is assumed in order to obtain robust results. Process capability analysis, consists one of those procedures that take for granted the normality assumption and thus, in order to achieve reliability and robust results, suitable transformations have to be implemented.

2.1 The Box-Cox Transformation

One way for transforming data into normally distributed is by applying the Box-Cox transformation [4]. In the Box-Cox transformation, there exists an exponential parameter λ which plays a key role and its values vary from -5 to 5. The optimal value for λ is the one that gives the best approach to a normal distribution curve.

Tukey in [36], proposed a transformation family where the transformed values are a monotonic function of the observations over some allowable range and is given by

$$y_i^{(\lambda)} = \begin{cases} y_i^{\lambda}, & \text{if } \lambda \neq 0 \\ \log y_i, & \text{if } \lambda = 0 \end{cases}, \quad y_i > 0, \tag{2.1}$$

for $y_i > 0$ where y_i are the values of the response variable. The problem of equation (2.1) is that: (1) the order is not preserved and; (2) there is discontinuity in the case of $\lambda = 0$.

Box and Cox in [4] solved the aforementioned problems by subtracting one and dividing with λ the transformation y_i^{λ}. The original Box-Cox transformation applies with parametric families of transformations and takes the following form:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log y_i, & \text{if } \lambda = 0 \end{cases}. \tag{2.2}$$

The transformation in the above equation (2.2) applies only for strictly positive numbers ($y_i > 0 \quad \text{for} \quad i = 1, ..., n$). The upper leg of the 2.2 is a scaled version of the Tukey transformation y_i^{λ}. As for the lower leg, when $\lambda \to 0$, $y_i^{(\lambda)} \to \log y_i$. Note

that for the 0^{th} power $y_i^0 = 1$, there is no meaning and the formula is rewritten as follows:

$$y_i^{(\lambda)} = \frac{e^{\lambda \log y_i} - 1}{\lambda} \approx \frac{\left(1 + \lambda \log(y_i) + \frac{1}{2}\lambda^2 \log(y_i)^2 + ...\right) - 1}{\lambda} \to \log(y_i), \qquad (2.3)$$

as $\lambda \to 0$. This result can also be achieved by using l'Hôpital's rule.

The fact that equation (2.2) can be applied only when $y_i > 0$, led Box and Cox to develop a second transformation, known as *"two-parameter Box-Cox transformation"*, which can also be used for negative values ($y_i < 0$).

Thus, the form of the transformation is as follows:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{(y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1}, & \text{if } \lambda_1 \neq 0 \\ \log(y_i + \lambda_2), & \text{if } \lambda_1 = 0 \end{cases} . \qquad (2.4)$$

In this case $\lambda = (\lambda_1, \lambda_2)'$, where $y_i > -\lambda_2$ or $y_i + \lambda_2 > 0$. The parameter λ_2 is set manually and is calculated for all y_i so that $y_i + \lambda_2$ is always positive.

2.1.1 Estimation of the Transformation Parameter

For the estimation of the parameter λ, Box and Cox considered two approaches. The first approach is the use of the **Maximum Likelihood method**. It is the most commonly used method since its easy to calculate the profile likelihood function and obtain approximate confident intervals for λ due to the asymptotic properties of the ML method. The second approach is based on the use of **Bayesian method**.

2.1.1.1 Maximum Likelihood Method

Consider that there is an $n \times 1$ vector of observations $\underset{\sim}{y} = \{y_1, y_2, ..., y_n\}$ and for an unknown λ

$$\underset{\sim}{y}^{(\lambda)} = \left\{y_1^\lambda, y_2^\lambda, ..., y_n^\lambda\right\} = \mathrm{A}\underset{\sim}{\theta} + \underset{\sim}{\varepsilon}, \qquad (2.5)$$

where $\underset{\sim}{y}^{(\lambda)}$ is a vector which contains the transformed observations, A is a known $n \times k$ matrix of constants, $\underset{\sim}{\theta}$ is a $k \times 1$ vector which contains unknown parameters related with the transformed values and $\underset{\sim}{\varepsilon} \sim N_n(\underset{\sim}{0}, \sigma^2 I_n)$ is the residuals vector. By calculating the mean of equation (2.5) we have

$$E(\underset{\sim}{y}^{(\lambda)}) = \mathrm{A}\underset{\sim}{\theta}. \qquad (2.6)$$

In this case, it can be presumed that the transformed observations $\underset{\sim}{y}^{(\lambda)} \sim N_n(\mathrm{A}\underset{\sim}{\theta}, \sigma^2 I_n)$ satisfy the assumption of normality and the model parameters are $(\lambda, \underset{\sim}{\theta}, \sigma^2)$. The probability density function (pdf) of the transformed observations is obtained by the formula:

$$f(\underset{\sim}{y}^{(\lambda)}) = \frac{1}{(2\pi)^{\frac{1}{2}n}\sigma^n} \exp\left\{-\frac{(\underset{\sim}{y}^{(\lambda)} - \mathrm{A}\underset{\sim}{\theta})'(\underset{\sim}{y}^{(\lambda)} - \mathrm{A}\underset{\sim}{\theta})}{2\sigma^2}\right\}. \qquad (2.7)$$

By multiplying equation (2.7) with the Jacobian determinant of the transformed values $\underset{\sim}{y}^\lambda$, the pdf of $\underset{\sim}{y}$ can be calculated, i.e., the likelihood for the whole model:

$$L(\lambda, \underset{\sim}{\theta}, \sigma^2 | \underset{\sim}{y}, \mathrm{A}) = f(\underset{\sim}{y}) = \frac{1}{(2\pi)^{\frac{1}{2}n}\sigma^n} \exp\left\{ -\frac{(\underset{\sim}{y}^{(\lambda)} - \mathrm{A}\underset{\sim}{\theta})'(\underset{\sim}{y}^{(\lambda)} - \mathrm{A}\underset{\sim}{\theta})}{2\sigma^2} \right\} J(\lambda; \underset{\sim}{y}), \quad (2.8)$$

where

$$J(\lambda; \underset{\sim}{y}) = \prod_{i=1}^{n} \left| \frac{dy_i^{(\lambda)}}{dy_i} \right|.$$

For each fixed λ, equation (2.8) results in the maximum likelihood estimators for $(\underset{\sim}{\theta}, \sigma^2)$

$$\widehat{\underset{\sim}{\theta}}(\lambda) = (\mathrm{A}'\mathrm{A})^{-1}\mathrm{A}\underset{\sim}{y}^{(\lambda)}, \quad (2.9)$$

and

$$\widehat{\sigma}^2(\lambda) = \frac{\underset{\sim}{y}^{(\lambda)'}\mathrm{A}_r\underset{\sim}{y}^{(\lambda)}}{n} = \frac{S(\lambda)}{n}, \quad (2.10)$$

where $S(\lambda)$ denotes the residual sum of squares of $\underset{\sim}{y}^{(\lambda)}$. When A is a full rank matrix, then

$$\mathrm{A}_r = I - \mathrm{A}(\mathrm{A}'\mathrm{A})^{-1}\mathrm{A}'. \quad (2.11)$$

By substituting $\widehat{\sigma}^2(\lambda)$ and $\widehat{\underset{\sim}{\theta}}(\lambda)$ into equation (2.8), the maximized log likelihood function (i.e., the profile log likelihood) for fixed λ can be written as:

$$L_{max}(\lambda) = L(\lambda | \underset{\sim}{y}, \mathrm{A}, \widehat{\underset{\sim}{\theta}}(\lambda), \widehat{\sigma}^2(\lambda)) = C - \frac{n}{2} log(\widehat{\sigma}^2(\lambda)) + \log J(\lambda; \underset{\sim}{y}), \quad (2.12)$$

where $C = \frac{n}{2}\log(2\pi/n) - \frac{n}{2}$.

If all the observations of the response variable are positive, according to equation (2.2) the term $\log J(\lambda; y)$ takes the form:

$$(\lambda - 1)\sum_{i=1}^{n} log(y_i). \quad (2.13)$$

However, if negatives values are observed, according to equation (2.4) the term $\log J(\lambda; y)$ transforms into:

$$(\lambda_1 - 1)\sum_{i=1}^{n} log(y_i + \lambda_2). \quad (2.14)$$

Note that the maximized log likelihood can be plotted against a series of values λ. From this plot the maximizing value of λ can be identified and an approximate $100(1-a)\%$ confidence region can be derived from:

$$L_{max}(\hat{\lambda}) - L_{max}(\lambda) < \frac{1}{2}\chi_{\nu_\lambda^2}(\alpha), \quad (2.15)$$

where ν_λ is the number of independent components in λ.

In case of different $\lambda's$, their comparison could be achieved by working with the normalized transformation:

$$\underset{\sim}{z}^{(\lambda)} = \underset{\sim}{y}^{(\lambda)}/J^{\frac{1}{n}}, \quad (2.16)$$

where $J = J(\lambda; y)$ is the Jacobian determinant of the transformation.

The profile log-likelihood can be simplified corresponding to equations (2.2) and (2.4) as:

$$\underset{\sim}{z}^{(\lambda)} = \begin{cases} \dfrac{\underset{\sim}{y}^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}, & \text{if } \lambda \neq 0 \\ \dot{y} \log \underset{\sim}{y}, & \text{if } \lambda = 0 \end{cases}, \tag{2.17}$$

while in the shifted location case, equation (2.17) can be written as:

$$\underset{\sim}{z}^{(\lambda)} = \begin{cases} \dfrac{(\underset{\sim}{y} + \lambda_2)^{\lambda_1} - 1}{\lambda_1 gm(\underset{\sim}{y} + \lambda_2)^{\lambda_1 - 1}}, & \text{if } \lambda_1 \neq 0 \\ gm(\underset{\sim}{y} + \lambda_2) \log(\underset{\sim}{y} + \lambda_2), & \text{if } \lambda_1 = 0 \end{cases}, \tag{2.18}$$

where, $\dot{y} = (\prod_{i=1}^{n} y_i)^{\frac{1}{n}}$ is the geometric mean of the observations and $gm(\underset{\sim}{y} + \lambda_2)$ is the geometric mean of the $(\underset{\sim}{y} + \lambda_2)$'s.

The profile log likelihood can be written as:

$$L_{max}(\lambda) = C - \frac{1}{2} n \log \widehat{\sigma}^2(\lambda; \underset{\sim}{z}), \tag{2.19}$$

where

$$\widehat{\sigma}^2(\lambda; \underset{\sim}{z}) = \frac{\underset{\sim}{z}^{(\lambda)'} \mathrm{A}_r \underset{\sim}{z}^{(\lambda)}}{n} = \frac{S(\lambda; \underset{\sim}{z})}{n}, \tag{2.20}$$

where $S(\lambda; \underset{\sim}{z})$ is the residuals sum of squares of $\underset{\sim}{z}^{(\lambda)}$. The maximum likelihood can be acquired by minimizing $S(\lambda; \underset{\sim}{z})$ with respect to λ.

2.1.1.2 Bayesian Method

An alternative way for parameter estimation can be achieved via Bayes's theorem. In order to obtain the posterior distribution of λ, the prior distributions of θ's and $\log \sigma$ must be considered uniformly distributed over the region where the likelihood is significant.

By using this approach the model must be recognizable and the design matrix must be of full column rank (i.e., each of the columns of the matrix are linearly independent) [28], otherwise θ's cannot be estimated. By rewriting the likelihood in equation (2.8) the conditional pdf of likelihood is:

$$p(\underset{\sim}{y} | \underset{\sim}{\theta}, \sigma^2, \lambda) = \frac{1}{(2\pi)^{\frac{1}{2}n} \sigma^n} \exp\left\{ - \frac{\nu_r s^2(\lambda) + (\underset{\sim}{\theta} - \widehat{\underset{\sim}{\theta}}_\lambda)' \mathrm{A}' \mathrm{A} (\underset{\sim}{\theta} - \widehat{\underset{\sim}{\theta}}_\lambda)}{2\sigma^2} \right\} J(\lambda; \underset{\sim}{y}), \tag{2.21}$$

where

$$s^2(\lambda) = \frac{\underset{\sim}{y}^{(\lambda)'} \mathrm{A}_r \underset{\sim}{y}^{(\lambda)}}{\nu_r} = \frac{S(\lambda)}{\nu_r}, \tag{2.22}$$

is the residual mean square of $\underset{\sim}{y}^{(\lambda)}$, $\nu_r = n - rank(\mathrm{A})$ are the degrees of freedom of the residuals and $\widehat{\underset{\sim}{\theta}}_\lambda$ is the least squares estimation of $\underset{\sim}{\theta}$ for a given λ.

The $\underset{\sim}{\theta}$'s are parameterized so that they are linearly independent and have $n - \nu_r$ degrees of freedom. Suppose that $p_0(\lambda)$ represents the marginal prior density of λ. For given λ the conditional prior distribution is

$$g(\lambda) d\underset{\sim}{\theta}_\lambda d(\log \sigma_\lambda). \tag{2.23}$$

The term $g(\lambda)$ indicates the existence of dependence between λ and the range of the values of $\underset{\sim}{y}^{(\lambda)}$. It can be specified by fixing a value of λ, say λ_1 and let us suppose that there exists a linear relationship between $\underset{\sim}{y}^{(\lambda)}$ and $\underset{\sim}{y}^{(\lambda_1)}$

$$\underset{\sim}{y}^{(\lambda)} = const + l_\lambda \underset{\sim}{y}^{(\lambda_1)}. \tag{2.24}$$

For every $g(\lambda)$ that corresponds to equation (2.24), the conditional prior distributions from equation (2.23) appears to be consistent for different values of λ. From equation (2.24) the following expression can be derived:

$$\log \sigma_\lambda^2 = const + \log \sigma_{\lambda_1}^2, \tag{2.25}$$

where σ_λ^2 is independent of λ. However,

$$\frac{d\underset{\sim}{\theta}_\lambda}{d\underset{\sim}{\theta}_{\lambda_1}} = l_\lambda,$$

where

$$l_\lambda = \{J(\lambda; \underset{\sim}{y})\}^{1/n}. \tag{2.26}$$

Therefore, the conditional prior density is given by

$$\frac{d\underset{\sim}{\theta}_\lambda d(\log \sigma_\lambda)}{\{J(\lambda; \underset{\sim}{y})\}^{(n-\nu_r)/n}}, \tag{2.27}$$

and by combining it with the marginal prior density of λ, the final form of equation (2.27) is obtained

$$\frac{d\underset{\sim}{\theta} d(\log \sigma)}{\{J(\lambda; \underset{\sim}{y})\}^{(n-\nu_r)/n}} p_0(\lambda) d\lambda. \tag{2.28}$$

With the help of the likelihood (equation (2.21)) and the prior density (equation (2.28)) the marginal posterior distribution can be estimated by applying the Bayes's theorem and hence, the posterior takes the form

$$K'_y \frac{I(\lambda|\underset{\sim}{y}) p_0(\lambda)}{\{J(\lambda; \underset{\sim}{y})\}^{(n-\nu_r)/n}}, \tag{2.29}$$

where K'_y denotes a normalizing constant which is independent from λ and

$$I(\lambda|\underset{\sim}{y}) = \int_{-\infty}^{\infty} d(\log \sigma) \int_{-\infty}^{\infty} d\underset{\sim}{\theta}\, p(\underset{\sim}{y}|\underset{\sim}{\theta}, \sigma^2, \lambda). \tag{2.30}$$

By calculating equation (2.30) the posterior distribution in equation (2.29) can be rewritten as

$$K'_y \frac{J(\lambda; \underset{\sim}{y})^{\nu_r/n}}{\{s^2(\lambda)\}^{\frac{1}{2}\nu_r}} p_0(\lambda). \tag{2.31}$$

The quantity,

$$\frac{J(\lambda; \underset{\sim}{y})^{\nu_r/n}}{\{s^2(\lambda)\}^{\frac{1}{2}\nu_r}}, \tag{2.32}$$

represents the contribution of the values $\underset{\sim}{y}$ to the posterior distribution of λ and on a log scale

$$L_b(\lambda) = -\frac{1}{2}\nu_r \log s^2(\lambda) + (\nu_r/n) \log J(\lambda; \underset{\sim}{y}). \tag{2.33}$$

If we consider the normalized transformation $\underset{\sim}{z}^{(\lambda)} = \underset{\sim}{y}^{(\lambda)}/J^{\frac{1}{n}}$, the above equation (2.33) can be rewritten as:

$$L_b(\lambda) = -\frac{1}{2}\nu_r \log s^2(\lambda; \underset{\sim}{z}) = -\frac{1}{2}\nu_r \log \{S(\lambda; \underset{\sim}{z})/\nu_r\}. \tag{2.34}$$

Observe that the difference between the log scale of the contribution to the posterior and the maximum likelihood in equation (2.19) is the replacement of n by ν_r.

Table 2.1 provides a list of some common transformations of the variable y_i^λ for specific λ values.

Table 2.1: Common Box-Cox Transformations

λ	Transformed data
-3	$y_i^{-3} = 1/y_i^3$
-2	$y_i^{-2} = 1/y_i^2$
-1	$y_i^{-1} = 1/y_i^1$
-0.5	$y_i^{-0.5} = 1/(\sqrt{y_i})$
0	$\log(y_i)$
0.5	$y_i^{0.5} = \sqrt{(y_i)}$
1	$y_i^1 = y_i$
2	y_i^2
3	y_i^3

2.2 The Yeo-Johnson Transformation

Besides the Box-Cox transformation for negative values (equation (2.4)), a number of alternative transformation families for y_i have been suggested that can handle the problem of non-positivity.

Among them, Yeo and Johnson in [37], suggested a transformation family that negative values are allowed without restrictions on y_i and have adapted many of the properties of the Box-Cox family:

$$\psi(\lambda, \underset{\sim}{y}) = y_i^\lambda = \begin{cases} ((y_i+1)^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, y_i \geq 0 \\ \log(y_i+1), & \text{if } \lambda = 0, y_i \geq 0 \\ -[(-y_i+1)^{2-\lambda} - 1]/(2-\lambda), & \text{if } \lambda \neq 2, y_i < 0 \\ -\log(-y_i+1), & \text{if } \lambda = 2, y_i < 0 \end{cases}. \tag{2.35}$$

Notice that when $\lambda \neq 0$ and $y_i \geq 0$, the transformation of $(y_i + 1)$ is the same as the one of Box-Cox, while when $\lambda \neq 2$ and $y_i < 0$, the transformation of $(-y + 1)$ is once again the Box-Cox transformation with power $(2 - \lambda)$. In case that both negative and positive values are present, then the transformation is calculated by using both types of transformations.

In the case of Yeo and Johnson, the estimated transformation parameter λ, minimizes the distance of Kullback-Leibler measure between the transformed values and the normal distribution.

2.3 Alternative Box-Cox Transformations

Manly in [23], proposed an alternative version of the Box-Cox transformation which can be applied for negative observations and it is considered a preferable choice when transforming skewed unimodal distributions into symmetric normal-like distributions and is given by

$$y_i^{(\lambda)} = \begin{cases} \dfrac{e^{(\lambda y_i)} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ y_i, & \text{if } \lambda = 0 \end{cases}. \tag{2.36}$$

John and Draper in [17], suggested a modification of the Box-Cox transformation known as *"Modulus Transformation"*. This transformation can also handle negative values and works better at normalizing distributions that have a bit of symmetry. It can be written as follows:

$$y_i^{(\lambda)} = \begin{cases} sign(y_i)\frac{(|y_i|+1)^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ sign(y_i)\log(|y_i|+1), & \text{if } \lambda = 0 \end{cases}. \tag{2.37}$$

Bickel and Doksum in [3], introduced another slightly different Box-Cox transformation that can include distributions of y_i^{λ} with unbounded support like the normal distribution. The formula they suggested is the following:

$$y_i^{(\lambda)} = \frac{\left|y_i^{\lambda}\right| sign(y_i) - 1}{\lambda}, \qquad \text{for } \lambda > 0, \tag{2.38}$$

where

$$sign(y_i) = \begin{cases} 1, & \text{if } y_i \geq 0 \\ -1, & \text{if } y_i < 0 \end{cases}.$$

Hawkins and Weisberg in [16], proposed a Box-Cox family with negative values allowed. The Hawkins-Weisberg family consists of a modification of the two parameter Box-Cox family, which allows negative responses. It uses the scaled power transformation of $y_i + \lambda_2$ in equation (2.4) and based on that calculates

$$z_i = 0.5\left(y_i + \left(y_i^2 + \lambda_2^2\right)^{1/2}\right), \quad \text{for } \; i = 1, ..., n. \tag{2.39}$$

The location parameter λ_2 is either estimated or manually selected. The quantity z is always positive for every possible value of y_i. The Box-Cox transformation is then applied to z_i's in (2.2) according to the following expression:

$$z_i^{(\lambda)} = \begin{cases} \dfrac{z_i^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log z_i, & \text{if } \lambda = 0 \end{cases}. \tag{2.40}$$

2.4 Box-Cox Transformation in Taguchi Analysis

The estimation of the optimum value of λ with the ordinary Box-Cox procedure by using maximum likelihood estimation, carries the risk of inconsistency of the error variance. In this section, we discuss the Taguchi procedure where the assumption of constant variance must be fulfilled and describe a method to achieve so.

2.4.1 Taguchi Procedure

Genichi Taguchi, a Japanese engineer and statistician, developed several methods for off-line quality control. The term off-line quality control refers to quality and cost control activities of a product. Taguchi by using design techniques such as robust design method, improved reliability and reduced product costs. That is he made the product procedure less sensitive to unexplained and uncontrolled variability factors (e.g., environmental variables) [19]. In statistical terms, Taguchi's purpose is to minimize the variability which is caused by unpredictable factors known as **noise**. The product in a robust design method is disturbed and controlled in general by the designer and noise factors. The noise factors can be controlled and simulated via controlled experimentation and an optimum combination of values can be derived that can handle the noise effects [35].

The results of the experimental trials are called **Performance Measures**. By analyzing the **Noise Performance Measures (NPM)**, useful results can be derived regarding the variability control factors and their optimal combination. Therefore, the analysis of **Target Performance Measures (TPM)**, which are connected with the process mean, can indicate which factors, except the ones of variability, have a significant effect on the mean response.

Figure 2.1, describes the robust design method. Notice that there are control and noise factors that effect the product or process. The variation between the response and the target value should be ideally the minimum possible.

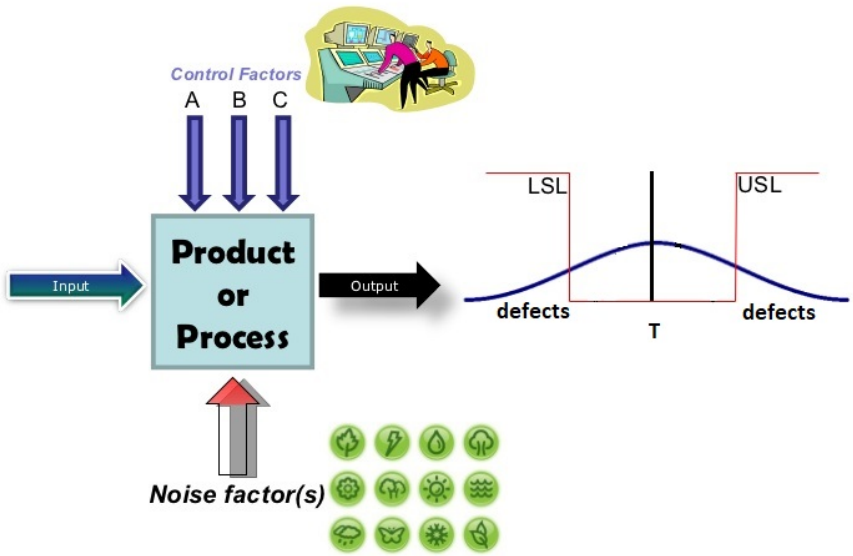

Figure 2.1: Aim of robust design

2.4.2 Independence, Secure and Proper Transformation

To obtain robust results in Taguchi analysis, independence between population mean and variance must occur. It can be achieved through proper transformation of the data with the condition of a well established functional relationship.

For instance, suppose that

$$\sigma_Y = f(\mu_Y) = a(\mu_Y)^\beta, \tag{2.41}$$

is a functional relationship between μ_Y and (σ_Y^2) of the raw data, and $T(Y)$ is a transformation for stabilizing the variance (i.e., σ_T^2 constant). By elaborating $T(Y)$ in a Taylor series (see [15]) the following equation is derived:

$$T(Y) = T(\mu_Y) + T'(\mu_Y)(Y - \mu_Y) + ... \quad . \tag{2.42}$$

By applying a variance operator in both sides of equation (2.42) and with the help of equation (2.41) we obtain:

$$T'(\mu_Y) = \sigma_T/\sigma_Y = \frac{\sigma_T}{a(\mu_Y)^\beta}, \tag{2.43}$$

where σ_T^2 is the variance of the transformed T(Y).

To make σ_T^2 constant, equation (2.43) must be integrated with respect to the mean such that:

$$T(\mu_Y) \approx \int \frac{c}{f(\mu_Y)} d\mu = \int \frac{\sigma_T}{a(\mu_Y)^\beta} d\mu. \tag{2.44}$$

The above approximation results in:

$$T(\mu_Y) \simeq \begin{cases} C_1 \mu_Y^{1-\beta}, & \text{if } \beta \neq 1 \\ C_1 \log \mu_Y, & \text{if } \beta = 1 \end{cases}, \tag{2.45}$$

where C_1 is a constant parameter. Suppose now that we want to apply the Box-Cox transformation. To ensure that the transformation can stabilize variance, λ must be approximately equal to:

$$\lambda = 1 - \beta. \tag{2.46}$$

By using equation (2.46) the transformation in equation (2.45) can be rewritten as:

$$T(\mu_Y) \simeq \begin{cases} C_1 \mu_Y^{\lambda}, & \text{if } \lambda \neq 0 \\ C_1 \log \mu_Y, & \text{if } \lambda = 0 \end{cases}. \tag{2.47}$$

Let us suppose that there are n samples from the same population, $\bar{x}_i$ denotes the sample mean and s_i denotes the sample variance for the $i-$th sample. By considering the logarithm on both sides of $\sigma_Y = a(\mu_Y)^\beta$, $\log a$ and β can be estimated through the least squares method.

$$\log s_i = \log a + \beta \log \bar{x}_i + \epsilon_i \quad \text{for} \quad i = 1, ..., n. \tag{2.48}$$

Equation (2.48) consists a linear regression model and thus, if the estimator of β holds approximately for $\lambda = 1 - \beta$, where λ is estimated by the regular Box-Cox maximum likelihood method, then the Box-Cox transformation stabilizes the variance.

In case that non-positive observations are present in the dataset, equation (2.41) can be expressed as:

$$\sigma \approx \alpha(\mu + \beta_1)^\beta, \tag{2.49}$$

where $y > -\beta_1$. Estimates of β and β_1 can be derived through

$$\log s_i = \log a + b \log(\bar{x}_i + b_1), \text{ for } i = 1, ..., n. \tag{2.50}$$

Considering that the transformed mean and variance are independent, proper performance measures can be established:

$$\begin{array}{l} \text{NPM} = -10 \log_{10} s_T^2 \\ \text{TPM} = \overline{y}_T \end{array} . \tag{2.51}$$

When the calculation of the integral in (2.44) in the original scale is hard to compute and the functional relationship between μ_Y and σ_Y^2 is known, the performance measures presented in equation (2.51) can be written as:

$$\begin{array}{l} \text{NPM} = 10 \log_{10} \left\{ f(\overline{y})/s \right\}^2 \\ \text{TPM} = \overline{y} \end{array} . \tag{2.52}$$

In case that equation (2.41) is applied and b denotes the estimator of β, the NPM of equation (2.52) can be rewritten as follows:

$$\text{NPM} = 10 \log_{10} \left(\frac{\overline{y}^b}{s} \right)^2 , \tag{2.53}$$

and in the special case of $b = 1$, equation (2.53) is equivalent to that of Taguchi's signal to noise ratio (S/N).
For a more detailed discussion on performance measures, the interested reader may refer to [21]; and [22].

Chapter 3

Analysis of Variance (ANOVA)

3.1 One-way ANOVA

Analysis of Variance has been proposed by Fisher (see in [11]) and it can be considered as an extension of the well known t-test, for the comparison of two means. Specifically it allows the comparison of k ($k > 2$) population means.

One-Way ANOVA, also called one-factor ANOVA, is a statistical test which investigates the equality of the population means. The term **one-way** points out that the populations are determined by only one explanatory variable (i.e., independent variable), usually referred to as **factor**.
The states of the explanatory variable are called **levels**, while the combination of factor levels that used to determine the response variable is called **treatment**. In One-Way ANOVA the factor can have at least two independent levels and each level can be viewed as a population. For example, suppose that we wish to study a population of children in high school and intent to examine the impact of three different teaching methods (A,B and C) in mathematics by their score on an arithmetic test. In this case, score is the dependent variable and teaching method is the factor which consists of three levels. Since there is only one factor, the means of the three populations will be compared based on one-way ANOVA.

3.1.1 Hypothesis Testing and Assumptions

Suppose that we want to compare k means $\mu_1, \mu_2, ..., \mu_k$. Thus, we are going to examine the following hypothesis:

$$\begin{gathered} H_0 : \mu_1 = \mu_2 = ... = \mu_k = 0 \\ \text{vs} \\ H_a : \text{At least one of } \mu_i \text{ is different } i = 1, 2, ..., k \end{gathered} \quad . \tag{3.1}$$

The null hypothesis (H_0) indicates that there exists equality between the groups, while the alternative hypothesis (H_a) indicates that there is no equality between the groups.

To make use of One-Way ANOVA, the following assumptions must be satisfied:

1. The **samples** associated with each level must be derived from a **normally distributed** population;

2. The **errors** must be **independent**;

3. The **variance of the populations** must be **equal** (homoskedasticity), and;

4. **The values of each sample** must be derived from **independent observations**.

3.1.2 One-Way ANOVA Model

Consider a single-factor variable with k treatments (populations) and n replicates (observations) for each level. Then the observations of an experiment for the One-Way ANOVA, could be modeled by:

$$y_{ij} = \mu_i + \epsilon_{ij}, \begin{cases} i = 1, 2, ..., k \\ j = 1, 2, ..., n \end{cases} , \tag{3.2}$$

where y_{ij} the j-th observation of the i-th level, μ_i the mean which corresponds to the i-th level (population) and ϵ_{ij} the random error of the j-th observation of the i-th level. By replacing the component μ_i with the following equation:

$$\mu_i = \mu + a_i, \quad i = 1, 2, ..., k \ ,$$

equation (3.2) can be rewritten as:

$$y_{ij} = \mu + a_i + \epsilon_{ij}, \begin{cases} i = 1, 2, ..., k \\ j = 1, 2, ..., n \end{cases} , \tag{3.3}$$

where μ is the total (overall) mean of all levels and a_i represents the effect of the i-th level and expresses the deviation between μ and the i-th level.

In the above setting, we assume that the number of observations per population is equal (n). It should be noted that as in the case of the standard t-test for two populations, the sample sizes could be different ($n_1, n_2, ..., n_k$).

One-Way ANOVA can be easily extended to the case where two or more factors are involved. In such a case, factors may act independently or interactively on the response variable. Thus, two settings should be considered: (1) Interaction between factors and; (2) No interaction between factors. The more general case of Two-Way ANOVA with interaction will be discussed in the following section.

3.2 Two−Way ANOVA

In order to gain reliable results for quality improvement, the majority of the experiments involve more than one factor. These kind of designs, which consist of two or more factors, are called **factorial designs**. The factorial designs contain in each replicate all the possible combinations of the factors levels. Based on the example presented in Section (3.1), we want to investigate if the three different teaching methods (A,B and C) and the age of each student (15,16,17 and 18 years old) affect their performance in an arithmetic test. In this case there are two factors: teaching methods which consists of three levels and age which consists of four levels. Each replicate of the experiment will include all 3×4 possible combinations of the factors levels and hence, Two-Way ANOVA will be used.

3.2.1 Two-Way ANOVA in Completely Randomized Design (CRD)

Generally, the randomization technique is used in designs in order to prevent systematic effects of noise factors. In a CRD, the treatments are randomly distributed in the experimental units. It is most commonly used in labs where uncontrollable factors are easy to control and the variability is exclusive derived from the noise factors.

Let us consider the general case of a two-factor experiment, where the first factor A has a levels ($i = 1, 2, ..., a$) and the second factor B has b levels ($j = 1, 2, ..., b$). The replicates have all equal sizes ($k = 1, 2, ..., n$) and thus, the design is considered balanced. The effects model, including interaction, can be written as follows:

$$y_{ijk} = \mu + a_i + \beta_j + (a\beta)_{ij} + \epsilon_{ijk}, \quad \begin{cases} i = 1, 2, ..., a \\ j = 1, 2, ..., b \\ k = 1, 2, ..., n \end{cases} , \tag{3.4}$$

where y_{ijk} is the $ijk-$th observation, μ is the overall mean, a_i is the $i-$th level effect of factor A, β_j is the $j-$th level effect of factor B, $(a\beta)_{ij}$ the interaction effect between the components a_i, β_j, and ϵ_{ijk} denotes the random error component, where $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$. The term $(a\beta)_{ij}$ describes whether the treatment mean differs from the additive model which is the same with the one in equation (3.4) excluding the interaction term.

3.2.1.1 Fixed Effects Model

The fixed effects model, reflects the situation where the data are collected from all levels of the factors of interest or all treatments conditions are contained in the study. In case that both factors have fixed effects, to avoid overparameterization in equation (3.4), constraints must be set to make the parameters interpretable. The following parameter restrictions for balanced designs are considered:

$$\sum_{i=1}^{a} a_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} (a\beta)_{ij} = \sum_{j=1}^{b} (a\beta)_{ij} = 0 \ . \tag{3.5}$$

Thus, the hypotheses testing for the case of Two-Way ANOVA by applying fixed effects in both factors are as follows:

$$\begin{aligned} &H_0 : a_1 = a_2 = ... = a_a = 0 \quad \text{vs} \quad H_1 : \text{at least one } \ a_i \neq 0 \\ &H_0 : \beta_1 = \beta_2 = ... = \beta_b = 0 \quad \text{vs} \quad H_1 : \text{at least one } \ \beta_j \neq 0 \ . \end{aligned} \tag{3.6}$$

Because there is an interaction term in the model presented in equation (3.4), the next hypothesis must be considered:

$$H_0 : (a\beta)_{ij} = 0 \ \ \forall \, i, j \quad \text{vs} \quad H_1 : \text{at least one } \ (a\beta)_{ij} \neq 0 \ . \tag{3.7}$$

For testing the above null hypotheses a plethora of calculations must be held.

The parameters of the model in (equation (3.4)) are estimated by:

$$\begin{aligned} \widehat{\mu} &= \overline{y}_{...} \\ \widehat{\alpha}_i &= \overline{y}_{i..} - \overline{y}_{...} \\ \widehat{\beta}_j &= \overline{y}_{.j.} - \overline{y}_{...} \\ \widehat{\alpha\beta}_{ij} &= \overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...} \end{aligned} \quad , \tag{3.8}$$

where $\overline{y}_{...}$ is the overall mean of all sample data, $\overline{y}_{i..}$ is the total number of all the mean responses under the $i-$th level of factor A, $\overline{y}_{.j.}$ denotes the total mean responses under the $j-$th level of factor B and $\overline{y}_{ij.}$ is the total number of all observations that corresponds to the $ij-$th cell. The quantity $\overline{y}_{ij.}$ is also known as **cell mean**.

Sum of Squares (SS) is a statistical technique which is used in ANOVA in order to measure the deviation of data points from the mean. The total variability of the observed response y_{ijk} can be calculated by the **Corrected Total Sum of Squares** (SS_T) and it can be expressed as follows:

$$\begin{aligned} SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{...})^2 =& \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}[(\overline{y}_{i..} - \overline{y}_{...}) + (\overline{y}_{.j.} - \overline{y}_{...}) \\ &+(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}) + (y_{ijk} - \overline{y}_{ij.})]^2 \\ =& bn\sum_{i=1}^{a}(\overline{y}_{i..} - \overline{y}_{...})^2 + an\sum_{j=1}^{b}(\overline{y}_{.j.} - \overline{y}_{...})^2 \\ &+n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 \\ &+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2 \ . \end{aligned} \tag{3.9}$$

An alternative way of writing equation (3.9) is by using the next abbreviations,

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E \ , \tag{3.10}$$

where

$$\begin{aligned} SS_A &= bn\sum_{i=1}^{a}(\overline{y}_{i..} - \overline{y}_{...})^2 \\ SS_B &= an\sum_{j=1}^{b}(\overline{y}_{.j.} - \overline{y}_{...})^2 \\ SS_{AB} &= n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 \\ SS_E &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2 \ . \end{aligned} \tag{3.11}$$

To that end, SS_T can be divided into the next individual components: (1) The component SS_A expressing the variation related solely to the level of factor A or the

the variability among the rows; (2) SS_B expressing the variation that is related solely to the level of factor B or the variability among columns; (3) SS_{AB} expressing the variation that occurs between the factors A and B or the variation between each row and column, and; (4) SS_E expressing the variation that occurs from the residuals among replicates.

Equation (3.9) or (3.10) is essential for the construction of the Two-Way ANOVA table. According to equation (3.5) the sum of the main effects for factors A and B, which have a and b levels respectively, is restricted to sum up to zero. This restriction indicates that the **degrees of freedom** (df) for factor A are $a-1$ and for factor B are $b-1$. The interaction effect also sum up to zero and by subtracting the df of the factors effects A and B from the df of the cells $(ab-1)$, we obtain the df of the interaction effect which are:

$$ab - 1 - (a-1) - (b-1) = (a-1)(b-1). \tag{3.12}$$

Within each cell of the AB levels there are $(n-1)$ df and hence, the df for the error term of all the cells will be $ab(n-1)$. By summing up the degrees of freedom for each SS in equation (3.10), the df for SS_T are derived as follows:

$$\begin{aligned} df(SS_T) &= df(SS_A) + df(SS_B) + df(SS_{AB}) + df(SS_E) \\ &= (a-1) + (b-1) + (a-1)(b-1) + ab(n-1) \\ &= abn - 1 \quad . \end{aligned} \tag{3.13}$$

By dividing SS_A, SS_B, SS_{AB} and SS_E with their corresponding df, the **mean squares** (MS) MS_A, MS_B and MS_{AB} are derived. The **expected** values of the aforementioned MS are given by the following equations:

$$\begin{aligned} E(MS_A) &= E\left(\frac{SS_A}{a-1}\right) = \sigma^2 + \frac{bn\sum_{i=1}^{a} a_i^2}{a-1} \\ E(MS_B) &= E\left(\frac{SS_B}{b-1}\right) = \sigma^2 + \frac{an\sum_{j=1}^{b} \beta_j^2}{b-1} \\ E(MS_{AB}) &= E\left(\frac{SS_{AB}}{(a-1)(b-1)}\right) = \sigma^2 + \frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}(a\beta)_{ij}^2}{(a-1)(b-1)} \\ E(MS_E) &= E\left(\frac{SS_E}{ab(n-1)}\right) = \sigma^2 \quad . \end{aligned} \tag{3.14}$$

The steps that must be followed in order to determine the expected MS, are presented in [26] and [13].

Under the null hypothesis the $F-$test statistic for each case is defined as follows:

$$\begin{aligned} H_0 : a_i = 0 \quad &, \quad F = \frac{MS_A}{MS_E} \overset{H_0}{\sim} F_{a-1,ab(n-1)} \\ H_0 : \beta_j = 0 \quad &, \quad F = \frac{MS_B}{MS_E} \overset{H_0}{\sim} F_{b-1,ab(n-1)} \quad . \\ H_0 : (a\beta)_{ij} = 0 \ &, \quad F = \frac{MS_{AB}}{MS_E} \overset{H_0}{\sim} F_{(a-1)(b-1),ab(n-1)} \end{aligned} \tag{3.15}$$

Based on (3.15) the form of the Two-Way ANOVA table is presented in Table 3.1.

Table 3.1: Two-Way ANOVA Table for Balanced CRD
(Fixed Effects for Both Factors)

SOURCE	**SS**	**df**	**MS**	**F**
Corrected Model	SS_R	$ab-1$	MS_R	MS_R/MS_E
Intercept	c	1	c	c/MS_E
A	SS_A	$a-1$	MS_A	MS_A/MS_E
B	SS_B	$b-1$	MS_B	MS_B/MS_E
A*B	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	MS_{AB}/MS_E
Error	SS_E	$ab(n-1)$	MS_E	-
Total	$c+SS_A+SS_B+SS_{AB}+SS_E$	abn	-	-
Corrected Total	SS_T	$abn-1$	-	-

The **Corrected Model** term with df $ab-1$, represents the overall model. This model includes the variation of the two factors separately as well as the variation of the interaction between the two factors. It does not include the variation of the residuals. When the experiment is balanced, the SS of the corrected model is denoted by:

$$\begin{aligned} SS_R &= SS_A + SS_B + SS_{AB} \\ &\text{or} \\ SS_R &= SS_T - SS_E \end{aligned} \quad . \tag{3.16}$$

Note that the intercept term in Table 3.1 is used to test if the total mean differs from zero.

3.2.1.2 Random Effects Model

The random effects model, is used in the case that the experimenter makes use of **sampled levels of factors from a larger population of factors** with the goal of deriving conclusions about the population in each level. To that end, not all the treatments of interest are included, but a random sample of them is used in the experiment. The response observations can be described in a same manner as in the linear model in equation (3.4). Let us consider the case of both factors effects being random. The model parameters a_i, β_j, $(a\beta)_{ij}$ and ϵ_{ijk} are independent random variables and normally distributed with zero mean and different variances. Thus, the assumptions for the random effects model are:

$$\begin{aligned} a_i &\overset{iid}{\sim} N(0,\sigma_a^2) \\ \beta_j &\overset{iid}{\sim} N(0,\sigma_\beta^2) \\ (a\beta)_{ij} &\overset{iid}{\sim} N(0,\sigma_{a\beta}^2) \\ \epsilon_{ijk} &\overset{iid}{\sim} N(0,\sigma^2) \end{aligned} \quad , \tag{3.17}$$

where σ_a^2, σ_β^2, $\sigma_{a\beta}^2$ and σ_ϵ^2 are called **variance components**. The sum of the variance components reenacts the variance for each response observation at the $ijk-$th level.

Therefore, the variance of y_{ijk} is expressed as:

$$Var(y_{ijk}) = \sigma_y^2 = \sigma_a^2 + \sigma_\beta^2 + \sigma_{a\beta}^2 + \sigma^2 \quad , \tag{3.18}$$

which indicates that every effect has different variance.

The hypotheses testing for Two-Way ANOVA by applying random effects in both factors will be the following:

$$\begin{aligned} H_0 : \sigma_a^2 = 0 \quad \text{vs} \quad H_1 : \sigma_a^2 > 0 \\ H_0 : \sigma_\beta^2 = 0 \quad \text{vs} \quad H_1 : \sigma_\beta^2 > 0 \\ H_0 : \sigma_{a\beta}^2 = 0 \quad \text{vs} \quad H_1 : \sigma_{a\beta}^2 > 0 \end{aligned} \quad . \tag{3.19}$$

The formulas to obtain SS, df and MS are the same with the ones used in the fixed effects model. However, to build the F-statistics for testing the above hypotheses, the expected MS need to be examined based on the following formulas (see in [26]):

$$\begin{aligned} E(MS_A) &= \sigma^2 + n\sigma_{a\beta}^2 + bn\sigma^2 a \\ E(MS_B) &= \sigma^2 + \sigma_{a\beta}^2 + an\sigma_\beta^2 \\ E(MS_{AB}) &= \sigma^2 + n\sigma_{a\beta}^2 \\ E(MS_E) &= \sigma^2 \end{aligned} \quad . \tag{3.20}$$

For example, for testing H_0 by using the $F-$ratio, one has to compare two expected MS and examine if under the H_0 are equal. Thus, in the $F-$ratio the MS with the biggest expected value will be the numerator while the MS with the lower expected value will be the denominator. Hence, $F-$tests under H_0 in equation (3.19), will be the following:

$$\begin{aligned} H_0 : \sigma_a^2 = 0, \quad F &= \frac{MS_A}{MS_{AB}} \overset{H_0}{\sim} F_{a-1,(a-1)(b-1)} \\ H_0 : \sigma_\beta^2 = 0, \quad F &= \frac{MS_B}{MS_{AB}} \overset{H_0}{\sim} F_{b-1,(a-1)(b-1)} \\ H_0 : \sigma_{a\beta}^2 = 0, \quad F &= \frac{MS_{AB}}{MS_E} \overset{H_0}{\sim} F_{(a-1)(b-1),ab(n-1)} \end{aligned} \quad . \tag{3.21}$$

Table 3.2, presents the Two-Way ANOVA table for random effects.

Table 3.2: Two-Way ANOVA Table for Balanced CRD
(Random Effects for Both Factors)

SOURCE	**SS**	**df**	**MS**	**F**
Corrected Model	SS_R	$ab-1$	MS_R	MS_R/MS_E
Intercept	c	1	c	c/MS_E
A	SS_A	$a-1$	MS_A	MS_A/MS_{AB}
B	SS_B	$b-1$	MS_B	MS_B/MS_{AB}
A*B	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	MS_{AB}/MS_E
Error	SS_E	$ab(n-1)$	MS_E	-
Total	$c + SS_A + SS_B + SS_{AB} + SS_E$	abn	-	-
Corrected Total	SS_T	$abn-1$	-	-

3.2.1.3 Mixed Effects Model

In a **mixed effects model**, the effect of the one factor is fixed and of the other is random.

Let us consider that A is a fixed factor and B is a random factor. We will consider two approaches about this setting. Suppose a table with a rows from factor A and b columns are selected at random from factor B. In this case, if the experiment is repeated and the same column is derived twice, then the same column will be projected in both experiments. As a result, the main effects of column will specify the differences between the columns and thus, the sum of the interactions effect in a column will equal zero. Because of this restriction the model is called **restricted model**.

In the second approach, a sample of columns effects regarding factor B, is chosen randomly and independently from the whole column population. The same one goes for the interactions effects which also are chosen randomly and independently from the population interaction effects. This approach is referred to as **unrestricted model** and is mainly used in unbalanced designs because of its ease of use in that specific occasion. For more information about unrestricted models, the interested reader may refer to [26].

In this Section the case of the restricted model will be discussed. The model which describes it, is the same as in equation (3.4), where a_i is a fixed effect for factor A and β_j is a random effect for factor B or vice-versa and ϵ_{ijk} is the random error term.
Thus, the restricted model is based on the following assumptions:

$$\begin{array}{c} \beta_j \overset{iid}{\sim} N(0, \sigma_\beta^2) \\ \epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \\ \sum_{i=1}^{\alpha} a_i = 0 \end{array} . \tag{3.22}$$

In addition, the interaction term $(a\beta)_{ij}$ is a random effect from normal distribution with zero mean and variance $[(a-1)/a]\sigma_{a\beta}^2$ (instead of the regular $\sigma_{a\beta}^2$ in order to make the form of the expected mean squares simpler and easier to describe). The sum of the interaction term over the fixed factor A will be equal to zero.

$$\sum_{i=1}^{\alpha} (a\beta)_{ij} = (a\beta)_{.j} = 0. \tag{3.23}$$

The restriction in (3.23) indicates that some interaction effects over different levels of the fixed factor ($i \neq i'$) will be dependent to each other and thus, the covariance of the interaction effects for $i \neq i'$ will be

$$Cov[(a\beta)_{ij}(a\beta)_{i'j}] = -\frac{1}{a}\sigma_{a\beta}^2, \tag{3.24}$$

while the covariance of the interaction effects for $j \neq j'$ will be equal to zero. The hypotheses testing for the restricted model is as follows:

$$\begin{array}{l} H_0 : a_i = 0 \quad vs \quad H_1 : \text{at least one } a_i \neq 0, \quad i = 1, ..., a \\ H_0 : \sigma_\beta^2 = 0 \quad vs \quad H_1 : \sigma_\beta^2 > 0 \\ H_0 : \sigma_{a\beta}^2 = 0 \quad vs \quad H_1 : \sigma_{a\beta}^2 > 0 \end{array} . \tag{3.25}$$

Before defining proper test statistics, the expected MS are defined as follows (see in [26]):

$$\begin{aligned} E(MS_A) &= \sigma^2 + n\sigma^2_{a\beta} + \frac{bn\sum_{i=1}^{\alpha} a_i^2}{\alpha - 1} \\ E(MS_B) &= \sigma^2 + an\sigma^2_{\beta} \\ E(MS_{AB}) &= \sigma^2 + n\sigma^2_{a\beta} \\ E(MS_E) &= \sigma^2 \quad . \end{aligned} \tag{3.26}$$

Thus, based on (3.26) the appropriate $F-$tests under H_0 in (3.25) will be the following:

$$\begin{aligned} H_0 : a_i = 0\,, \quad F &= \frac{MS_A}{MS_{AB}} \overset{H_0}{\sim} F_{a-1,(a-1)(b-1)} \\ H_0 : \sigma^2_{\beta} = 0\,, \quad F &= \frac{MS_B}{MS_E} \overset{H_0}{\sim} F_{b-1,ab(n-1)} \quad . \\ H_0 : \sigma^2_{a\beta} = 0, \quad F &= \frac{MS_{AB}}{MS_E} \overset{H_0}{\sim} F_{(a-1)(b-1),ab(n-1)} \end{aligned} \tag{3.27}$$

Notice that the MS of the fixed factor A is divided by the MS of the interaction AB while the rest of the factors, which are random, are divided by the MS of the errors.

Finally, the Two-Way ANOVA table for mixed effects (A fixed, B random) is presented in Table 3.3 .

Table 3.3: Two-Way ANOVA Table for Balanced CRD
(Mixed Effects: A Fixed Effect and B Random Effect)

SOURCE	**SS**	**df**	**MS**	**F**
Corrected Model	SS_R	$ab-1$	MS_R	MS_R/MS_E
Intercept	c	1	c	c/MS_E
A	SS_A	$a-1$	MS_A	MS_A/MS_{AB}
B	SS_B	$b-1$	MS_B	MS_B/MS_E
A*B	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	MS_{AB}/MS_E
Error	SS_E	$ab(n-1)$	MS_E	-
Total	$c + SS_A + SS_B + SS_{AB} + SS_E$	abn	-	-
Corrected Total	SS_T	$abn-1$	-	-

Consider now the case where the factor A has random effects and the factor B has fixed. By using the linear model in equation (3.4), a_i is a random effect for A, β_j is a fixed effect for B, $(a\beta)_{ij}$ is a random effect and ϵ_{ijk} is the random error. The assumptions in this case are the following:

$$\begin{aligned} a_i &\overset{iid}{\sim} N(0, \sigma^2_a) \\ (a\beta)_{ij} &\overset{iid}{\sim} N(0, \sigma^2_{a\beta}) \\ \epsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma^2) \\ \sum_{j=1}^{b} \beta_j &= 0 \quad . \\ \sum_{j} a\beta &= 0 \end{aligned} \tag{3.28}$$

and thus, the hypotheses for testing are:

$$\begin{array}{lll} H_0 : \sigma_a^2 = 0 & \text{vs} & H_1 : \sigma_a^2 > 0 \\ H_0 : \beta_j = 0 & \text{vs} & H_1 : \text{at least one } \beta_j \neq 0 \\ H_0 : \sigma_{a\beta}^2 = 0 & \text{vs} & H_1 : \sigma_{a\beta}^2 > 0 \end{array} \quad . \tag{3.29}$$

Once again, before defining proper test statistics, the expected MS must be defined (see in [29]):

$$\begin{aligned} E(MS_A) &= \sigma^2 + bn\sigma_a^2 \quad , \\ E(MS_B) &= \sigma^2 + n\sigma_{a\beta}^2 + \frac{an\sum_{j=1}^{b}\beta_j^2}{b-1} \quad . \\ E(MS_{AB}) &= \sigma^2 + n\sigma_{a\beta}^2 \\ E(MS_E) &= \sigma^2 \end{aligned} \tag{3.30}$$

Thus, based on (3.30) the appropriate $F-$tests under the H_0 presented in (3.29) will be:

$$\begin{array}{lll} H_0 : \sigma_a^2 = 0 \quad , & F = \dfrac{MS_A}{MS_E} \overset{H_0}{\sim} F_{a-1,ab(n-1)} & \\ H_0 : \beta_j = 0 & F = \dfrac{MS_B}{MS_{AB}} \overset{H_0}{\sim} F_{b-1,(a-1)(b-1)} & . \\ H_0 : \sigma_{a\beta}^2 = 0 & F = \dfrac{MS_{AB}}{MS_E} \overset{H_0}{\sim} F_{(a-1)(b-1),ab(n-1)} & \end{array} \tag{3.31}$$

Table 3.4, presents the Two-Way ANOVA table for mixed effects (A random, B fixed).

Table 3.4: Two-Way ANOVA Table for Balanced CRD
(Mixed Effects: A Random Effect and B Fixed Effect)

SOURCE	**SS**	**df**	**MS**	**F**
Corrected Model	SS_R	$ab-1$	MS_R	MS_R/MS_E
Intercept	c	1	c	c/MS_E
A	SS_A	$a-1$	MS_A	MS_A/MS_E
B	SS_B	$b-1$	MS_B	MS_B/MS_{AB}
A*B	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	MS_{AB}/MS_E
Error	SS_E	$ab(n-1)$	MS_E	-
Total	$c+SS_A+SS_B+SS_{AB}+SS_E$	abn	-	-
Corrected Total	SS_T	$abn-1$	-	-

3.2.2 Two-Way Repeated Measures ANOVA

In many aspects of statistical quality control, there exist differences between the subjects (i.e., experimental units) which are used in a design. For example, different people based on their experience or training, will produce different results in the response variable. In such a case, that difference will affect the experimental error. To that end, to test the equality of means we make use of a technique known as **Repeated Measures** (RM). A RM design (also known as within-subject design), is applied when each subject is exposed to each of the treatment conditions [27] and

thus, it can be considered that the measurements are repeated over time. Note that each treatment condition (or treatment level) represents a time point in which the subject is measured.
In the following Sections brief discussion on Two-Way ANOVA with RM is made.

3.2.2.1 Between and Within Subjects Variability

In the repeated measures ANOVA, due to the fact that every participant appears in every condition of the design, there are not individual differences in between subject variation. In contrast to One and Two-Way ANOVA, those differences are expressed through the error term and for such a reason, between subject variance is expressed as:

$$\textit{between subject variance} = \textit{treatment effect} + \textit{experimental error} \ .$$

In One Way ANOVA the variability within treatments conditions are composed by the individual differences and the experimental error. In RM ANOVA the within subject variability can be separated into: 1) variance resulting from individual differences and; (2) variance resulting from the experimental error. Due to the fact that, individuals are measured several times, the derived variance from individual differences can be excluded. Thus, the $F-$ratio without individual differences will be as follows:

$$F = \frac{\textit{treatment effect} + \textit{experimental error}}{\textit{experimental error}} \ .$$

Note that since the variation of individual differences is removed, the $F-$ratio of the repeated measures becomes more "powerful" than any type of ANOVA using between subjects designs.

3.2.2.2 Repeated Measures Linear Model

Lets us suppose that A denotes the between subject factor with a levels and B represents the within subject factor which is repeated with b levels. Also, a random sample consisted of n_i subjects are observed in factor A at $i-$th level. As a result, the classical Two-Way RM ANOVA on one factor design can be described as follows:

$$y_{ijk} = \mu + a_i + \beta_j + (a\beta)_{ij} + \pi_{k(i)} + (\beta\pi)_{jk(i)} + \epsilon_{ijk}, \quad \begin{cases} i = 1, 2, ..., a \\ j = 1, 2, ..., b \\ k = 1, 2, ..., n \end{cases} , \qquad (3.32)$$

where,

- a_i is the $i-$th group fixed effect of factor A ($\sum_{i=1}^{a} a_i = 0$);
- β_j is the $j-$th treatment fixed effect of factor B ($\sum_{j=1}^{b} \beta_j = 0$);
- $(a\beta)_{ij}$ denotes the fixed interaction effect between the $i-$th group and $j-$th treatment ($\sum_{i=1}^{a}(a\beta)_{ij} = \sum_{j=1}^{b}(a\beta)_{ij} = 0$);
- $\pi_{k(i)}$ are random effect parameters of the subject k nested within group i with $\pi_{k(i)} \sim N(0, \sigma_{\pi}^2)$;
- $(\beta\pi)_{jk(i)}$ is the random joint interaction effect of subject k and treatment j nested within group i with $(\beta\pi)_{jk(i)} \sim N(0, \frac{b-1}{b}\sigma_{\beta\pi}^2)$;

- ϵ_{ijk} is the random error term, which is assumed that $\epsilon_{ijk} \sim N(0, \sigma^2)$.

Thus, the statistical hypotheses for the model in (3.32) can be expressed as follows:

$$\begin{array}{lll} H_0 : a_1 = a_2 = ... = a_a = 0 & \text{vs} & H_1 : \text{at least one } a_i \neq 0 \\ H_0 : \beta_1 = \beta_2 = ... = \beta_b = 0 & \text{vs} & H_1 : \text{at least one } \beta_j \neq 0 \\ H_0 : (a\beta)_{11} = ... = (a\beta)_{ab} = 0 & \text{vs} & H_1 : \text{at least one } (a\beta)_{ij} \neq 0 \end{array} \quad . \tag{3.33}$$

The expected MS of this process will be the following:

$$\begin{aligned} E(MS_A) &= \sigma^2 + b\sigma_\pi^2 + \frac{b\sum_{i=1}^a n_i a_i^2}{a-1} \\ E(MS_B) &= \sigma^2 + \sigma_{\beta\pi}^2 + \frac{N\sum_{j=1}^b \beta_j^2}{b-1} \\ E(MS_{AB}) &= \sigma^2 + \sigma_{\beta\pi}^2 + \frac{\sum_{i=1}^a \sum_{j=1}^b n_i (a\beta)_{ij}^2}{(a-1)(b-1)} \\ E(MS_{B\times swg}) &= \sigma^2 + \sigma_{\beta\pi}^2 \\ E(MS_{swg}) &= \sigma^2 + b\sigma_\pi^2 \end{aligned} \quad , \tag{3.34}$$

where the swg denotes the term subject within variability. Thus, based on (3.34) the appropriate $F-$tests under H_0 will be the following:

$$\begin{aligned} H_0 : a_i = 0, \quad & F = \frac{MS_A}{MS_{swg}} \overset{H_0}{\sim} F_{a-1,N-a} \\ H_0 : \beta_i = 0, \quad & F = \frac{MS_B}{MS_{B\times swg}} \overset{H_0}{\sim} F_{b-1,(N-a)(b-1)} \\ H_0 : (a\beta)_{ij} = 0, \quad & F = \frac{MS_A B}{MS_{B\times swg}} \overset{H_0}{\sim} F_{(a-1)(b-1),(N-a)(b-1)} \end{aligned} \quad . \tag{3.35}$$

For $N = \sum_{i=1}^a n_i$ (unbalanced design) the ANOVA table for the RM model presented in (3.32) is given in Table 3.5.

Table 3.5: Two-Way RM ANOVA Table for Unbalanced Design (Repeated Measures on One Factor)

SOURCE	**df**	**SS**	**MS**	**F**
Between Subjects				
A	$a-1$	SS_A	MS_A	MS_A/MS_{swg}
Within Subjects (swg)	$N-a$	SS_{swg}	MS_{swg}	
Within Subjects				
B	$b-1$	SS_B	MS_B	$MS_B/MS_{B\times swg}$
AB	$(a-1)(b-1)$	SS_{AB}	MS_{AB}	$MS_{AB}/MS_{B\times swg}$
$B\times$ Subjects Within Groups ($B \times swg$)	$(N-a)(b-1)$	$SS_{B\times swg}$	$MS_{B\times swg}$	

In case of a balanced design, i.e., $n_1 = n_2 = ... = n_a = n$, Table 3.5 will be the same except quantity $N = na$ which will represent the total number of observations.

Let us suppose now the case where B and C are both within subject factors with b and c levels respectively and in each bc levels, a random sample of n people

is assigned. This kind of design is known as *Two-Way RM model with repeated measures in both factors* and the linear model that describes it is the following:

$$y_{jkl} = \mu + \beta_j + \gamma_k + (\beta\gamma)_{jk} + S_i + (\beta S)_{ji} + (\gamma S)_{ki} + (\beta\gamma S)_{jki} + \epsilon_{jki} \ , \begin{cases} i = 1, ..., n \\ j = 1, ..., b \\ k = 1, ..., c \end{cases} , \tag{3.36}$$

where

- β_j is the $j-$th fixed effect of the within subject factor B;
- γ_k is the $k-$th fixed effect of the within subject factor C;
- $(\beta\gamma)_{jk}$ is the B_jC_k interaction effect;
- S_i is the $i-$th subject random effect with $S_i \sim N(0, \sigma_i^2)$;
- $(\beta S)_{ji}$ denotes the interaction random effect of the $i-$th subject and factor B_j with $(\beta S)_{ji} \sim N(0, \sigma_{ji}^2)$;
- $(\gamma S)_{ki}$ denotes denotes the interaction random effect of the $i-$th subject and factor C_k with $(\gamma S)_{ki} \sim N(0, \sigma_{ki}^2)$;
- $(\beta\gamma S)_{jki}$ is the interaction random effect of the $i-$th subject and factors B_j and C_k with $(\beta\gamma S)_{jki} \sim N(0, \sigma_{jki}^2)$;
- ϵ_{jki} is the random error term with $\epsilon_{jki} \sim N(0, \sigma^2)$.

The ANOVA table for the RM model presented in (3.36) is presented in Table 3.6.

Table 3.6: Two-Way RM ANOVA Table for balanced Design
(Repeated Measures on Both Factors)

SOURCE	**df**	**SS**	**MS**	**F**
Between Subjects				
Subjects (S)	$n-1$	SS_S	MS_S	
Within Subjects				
B	$b-1$	SS_B	MS_B	$MS_B/MS_{B\times S}$
C	$(c-1)$	SS_C	MS_C	$F_C = MS_C/MS_{C\times S}$
$C \times S$	$(c-1)(n-1)$	$SS_{C\times S}$	$MS_{C\times S}$	
BC	$(b-1)(c-1)$	SS_{BC}	MS_{BC}	$F_{BC} = MS_{BC}/MS_{BC\times S}$
$BC \times S$	$(b-1)(c-1)(n-1)$	$SS_{BC\times S}$	$MS_{BC\times S}$	

For more information on RM designs the interested reader may refer to [1]; [14]; and [9].

Chapter 4

A New Transformation Approach for Time-Series Monitoring Data

As mentioned in Chapter 2, independence of sample mean and variance in each experimental trial, plays a key role for having sharply defined results in Taguchi analysis. In this Chapter, new adjusted transformation methods will be proposed and implemented. In addition, a comparison between the proposed transformations, the original Box-Cox transformation and the safeguard against mean bias method of Logothetis, will be held.

4.1 Introduction

Data transformations are widely used in many aspects of statistical quality control. By using transformations one can deal with violations of the assumptions of the response variable. The Box-Cox transformation [4] can also be used in a Taguchi analysis (off-line quality control) in order to establish suitable measures for the noise performance measures (see [34]). A plethora of transformations have been suggested over the years for transforming positive or negative values that appear in a dataset. Some of these remarkable transformations were created by Logothetis [21], Cook and Weisberg [8] and Yeo and Johnson [37]. Thus, selecting a proper transformation for the response variable, is very important in order to gain reliable results in statistical quality control.

There are some datasets in a Taguchi analysis though, where the properties of a Box-Cox transformation are valid except the constancy of the variance in each experimental unit. For example, when running a Box-Cox transformation, there exists the risk of oversimplification of the model which may result to a mean bias in error variance (i.e., there is no constancy of error variance).

To that end, Logothetis in [21, 22] presented a safeguard method which although ensures the independence between mean and variance in each trial, the model has some flaws concerning its accuracy and predictive power.

One of the disadvantages in a Taguchi analysis, is the vagueness in the classification of variability and target control values. Thus, straightforward results about which factors affect the variation and mean cannot be clearly derived. When this overlap in classification occurs, the assumption of constancy in error variance might not apply. In such a case, a proper transformation is needed which can result to independence between the mean and variance of each experimental trial. Logothetis

proposed the functional relationship $\sigma = f(\mu)$ between the sample mean and sample variance of the response variable for the original scale. By assuming the special case where $f(\mu) = a\mu^k$, the non-linear relationship that results between the sample mean and variance can be expressed as follows:

$$\sigma = a\mu^k. \tag{4.1}$$

In this case, the transformation that ensures constancy of the error variance can be described by equation (2.45). When n is the total number of the experimental trials, the power k can be estimated by converting equation (4.1) into a simple linear regression model between $\log\sigma$ and $\log\mu$:

$$\log\sigma = \log a + k\log\mu + \epsilon. \tag{4.2}$$

In SQC, and more specifically in a Taguchi analysis, noise performance measures are used in order to recognize the variability control factors. When no functional relationship exists between the mean and variance, the NPM can be expressed as:

$$\text{NPM} = -10\log_{10}(s_T^2), \tag{4.3}$$

where s_T^2 is the variance of the transformed data, while in case a functional relationship exists, like the one in equation (4.1), NPM is given by the following formula:

$$\text{NPM} = 10\log_{10}\left(\frac{f(\overline{x})}{\sigma}\right)^2. \tag{4.4}$$

As aforementioned, in Logothetis method by converting equation (2.45) into a simple linear model, k can be estimated. The flaw of this technique lies in the fact that the estimated value of k is calculated based on the linear model in equation (2.48) or equation (2.50), for positive and negative values, respectively, and not based on the non-linear model in equation (2.45). Thus, the analysis is based on the modeling of the logarithm of standard deviation and not on the actual one.

In this Chapter we introduce an alternative approach to that of Logothetis for model selection, which encounters the flaws of the latter. The proposed transformation satisfies the assumption of independence between the cell mean and variance of each experimental trial and thus, a quite accurate model can be achieved. Finally, the proposed approach is adequate for the case where negative values appear into the analysis, and hence, no further formulas are needed.

4.2 Methodology

4.2.1 The Simple Polynomial

In order to establish the relation between cell mean and variance of the response variable, we will make use of equation (4.1). The components a and k are being estimated through least squares and an exhaustive search of possible values, respectively. The resulted estimators are compared based on their mean squared error via ANOVA. The estimators with the minimum mean square error, will be analyzed and fitted in equation (4.1). This approach is similar to Akaike Criterion Information

(AIC), since all the compared models have the same penalty term. It is reminded that AIC is expressed as follows:

$$AIC = -2 \log Lik + pn, \tag{4.5}$$

where Lik denotes the likelihood, n is the sample size and p is the number of the parameters which included in each one of the candidate models.

4.2.2 The Full Polynomial

Except the simple polynomial of equation (4.1) we also suggest a full polynomial approach which is described by:

$$\sigma = a_0 + a_1\mu + a_2\mu^2 + ... + a_k\mu^k. \tag{4.6}$$

The degree of the polynomial, and thus the form of the model, is chosen by model selection techniques specifically via AIC, MSE, and R^2.

Note that the model mentioned in equation (4.6), is appropriate for both positive and negative values into the dataset. Observing the models proposed by Box-Cox (equation (2.2)) and Logothetis (equation (2.41)), it is clear that both could be considered as sub-cases of the one in equation (4.6).

According to the above methods concerning the polynomial models, the mean and variability of a process can be expressed and monitored through adjusted control charts such as $\overline{X}-$chart, $S-$chart, etc.. These charts are considered beneficial in gaining information about the state of a process. That is, we can determine if the process is IC or changes are observed that shift the process in an OOC state. In the following section, the proposed transformations are being implemented in both real and simulated data.

In Section 4.3 we analyze a real dataset as well as simulated data, to explore the capability of the two proposed methods.

4.3 Case Studies

4.3.1 Real Case

The dataset presented in Table 4.1, represents a Taguchi's Orthogonal Array design $OA_{18}3^6$ (see in [34]) and consists of eighteen (18) number of tests on each variable (i.e., # of trials) and three (3) measurements for each combination of six (6) factors with three (3) levels each.

Table 4.1: Taguchi's Orthogonal Array ($OA_{18}3^6$)

INTERNAL ORDER-CONTROL FACTORS									
TRIAL	1	2	3	4	5	6	DATA		
	A	B	C	D	E	F	Y1	Y2	Y3
1	1	1	1	1	1	1	10.4	10.6	10.8
2	2	2	2	2	2	2	9.8	9.9	9.7
3	3	3	3	3	3	3	9.1	9.1	9.2
4	1	1	2	2	3	3	10.2	10.3	10.5
5	2	2	3	3	1	1	9.5	9.6	9.7
6	3	3	1	1	2	2	9.1	9	8.9
7	1	2	1	3	2	3	9.9	9.6	9.5
8	2	3	2	1	3	1	9.2	9.3	9.1
9	3	1	3	2	1	2	9.3	9.4	9.5
10	1	3	3	2	2	1	9.4	9.5	9
11	2	1	1	3	3	2	10	10.3	9.9
12	3	2	2	1	1	3	9	9.2	9.1
13	1	2	3	1	3	2	9.8	9.6	9.9
14	2	3	1	2	1	3	9.2	9.1	9.5
15	3	1	2	3	2	1	9.3	9.2	9.3
16	1	3	2	3	1	2	9.2	9.1	9.4
17	2	1	3	1	2	3	10.5	10.4	10.7
18	3	2	1	2	3	1	9.5	9.4	9.6

By applying the:

(a) standard Taguchi performance measure;

(b) Logothetis measure;

(c) original Box-Cox transformation;

(d) simple polynomial approach referred to as KKLP (Kalligeris-Karagrigoriou-Ladopoulos-Parpoula);

(e) full polynomial approach referred to as Ladopoly (Ladopoulos polynomial).

to both mean and variance, the results presented in Tables 4.2−4.6 are obtained.

Table 4.2: *General Linear Model: Average versus A; B; C; D; E; F*

Source	**DF**	**Adj SS**	**Adj MS**	**F-Value**	**P-Value**
A	2	9.645	4.823	1813.840	**0.000**
B	2	0.612	0.306	115.030	**0.000**
C	2	0.032	0.016	5.960	0.013
D	2	0.517	0.026	9.720	0.002
E	2	1.004	0.502	188.790	0.000
F	2	0.011	0.005	1.990	0.174
Error	14	0.037	0.003		
Total	26	11.392			

Table 4.3: *General Linear Model: NPM (Taguchi) versus A; B; C; D; E; F*

Source	**DF**	**Adj SS**	**Adj MS**	**F-Value**	**P-Value**
A	2	0.979	0.490	6.86	**0.037**
B	2	1.698	0.849	11.88	**0.013**
C	2	0.095	0.047	0.66	0.555
D	2	0.088	0.044	0.62	0.576
E	2	0.046	0.023	0.32	0.74
F	2	0.037	0.019	0.26	0.78
Error	5	0.357	0.071		
Total	17	3.300			

Table 4.4: *General Linear Model: NPM (Logothetis) versus A; B; C; D; E; F*

Source	**DF**	**Adj SS**	**Adj MS**	**F-Value**	**P-Value**
A	2	146.039	73.019	10.24	**0.017**
B	2	2.189	1.094	0.15	0.862
C	2	43.208	21.604	3.03	0.137
D	2	11.316	5.658	0.79	0.502
E	2	4.379	2.190	0.31	0.749
F	2	2.800	1.400	0.2	0.828
Error	5	35.658	7.132		
Total	17	245.588			

Table 4.5: *General Linear Model: NPM (Box-Cox) versus A; B; C; D; E; F*

Source	**DF**	**Adj SS**	**Adj MS**	**F-Value**	**P-Value**
A	2	0.059	0.030	6.95	**0.036**
B	2	0.103	0.051	12.06	**0.012**
C	2	0.006	0.003	0.66	0.555
D	2	0.005	0.003	0.59	0.587
E	2	0.003	0.001	0.33	0.731
F	2	0.002	0.001	0.25	0.787
Error	5	0.021	0.004		
Total	17	0.199			

Table 4.6: *General Linear Model: NPM (KKLP) versus A; B; C; D; E; F*

Source	**DF**	**Adj SS**	**Adj MS**	**F-Value**	**P-Value**
A	2	88.317	44.158	4.56	**0.075**
B	2	12.942	6.471	0.67	**0.553**
C	2	33.252	16.626	1.72	0.271
D	2	8.570	4.285	0.44	0.665
E	2	7.279	3.639	0.38	0.705
F	2	2.119	1.059	0.11	0.899
Error	5	48.435	9.687		
Total	17	200.915			

Table 4.2 has been calculated based on the mean of the design while the rest of the (Tables 4.3 - 4.6) have been calculated based on the variability of the design. Tables 4.3 and 4.5, which correspond to the Taguchi and Box-Cox transformations, respectively, can both detect that factors A and B are affecting the variability of the process because of their small P-values ($P-$values< 0.05). Same conclusions can be derived for Table 4.2, since factors A and B have also small P-values ($P-$values$= 0.000$). In Logothetis method (Table 4.4), the problem of noise is partially solved since only factor A affects the variability ($P-$value< 0.05). The KKLP method (Table 4.6), seems to fully resolve the problem of noise, since neither of the factors A and B affect the variability of the experiment ($P-$values> 0.05).

4.3.2 Simulation

The simulated data were created based on the mean and standard deviation results of eight (8) factors according to the design $OA_{18}(2 \times 3^7)$ (Logothetis Table 4.2, p.103 [22]) with the use of R (see [12]). The sample statistics of mean and variance from the suggested techniques will be calculated and depicted on a $\overline{X} - S$ control chart [33] in order to evaluate the two basic characteristics of a procedure.

To achieve so we apply the:

(a) standard $\overline{X} - S$ control chart;

(b) Logothetis transformation given in equation (4.1), $\sigma = f(\mu) \equiv \sigma_1$ & $\mu = f^{-1}(\sigma) \equiv \mu_1$;

(c) Box-Cox transformation as presented in equation (2.2), $y_{bc} = \frac{y^{\lambda}-1}{\lambda-1}$;

(d) Ladopoly given in equation (4.1), $\sigma = f_{Ladopoly}(\mu) \equiv \sigma_3$;

(e) KKLP given in equation (4.6), $\sigma = f_{KKLP}(\mu) \equiv \sigma_2$ and $\mu = f^{-1}_{KKLP}(\sigma) \equiv \mu_2 + (\sigma) \equiv \mu_3$.

In Figure 4.1, $\overline{X}$, μ_1, μ_2 and μ_3, are represented in a multiple $\overline{X}-$Graph while in (4.2) S, σ_1, σ_2, σ_3 are represented in a multiple $S-$Graph. The sample which corresponds to the time point t_{14}, can be revealed OOC only from the Logothetis and KKLP methods while the other two fail to reveal it. Observe that the Ladopoly method is the closest one to $\overline{X}$ in terms of explaining it more accurate (i.e., there is no much difference between the lines). Ladopoly method is describing more accurately the observed values regressed on the S-values since it can detect the OOC point in t_{14}. Also, KKLP seems to be better than Logothetis method in terms of modeling more precisely the underline characteristic σ. Eventually, Logothetis and KKLP methods are comparable since they are both strong at detecting OOC points but it should be highlighted that Logothetis estimation methodology is based on the logarithmic transformation of σ in contrast to KKLP method which uses the raw data.

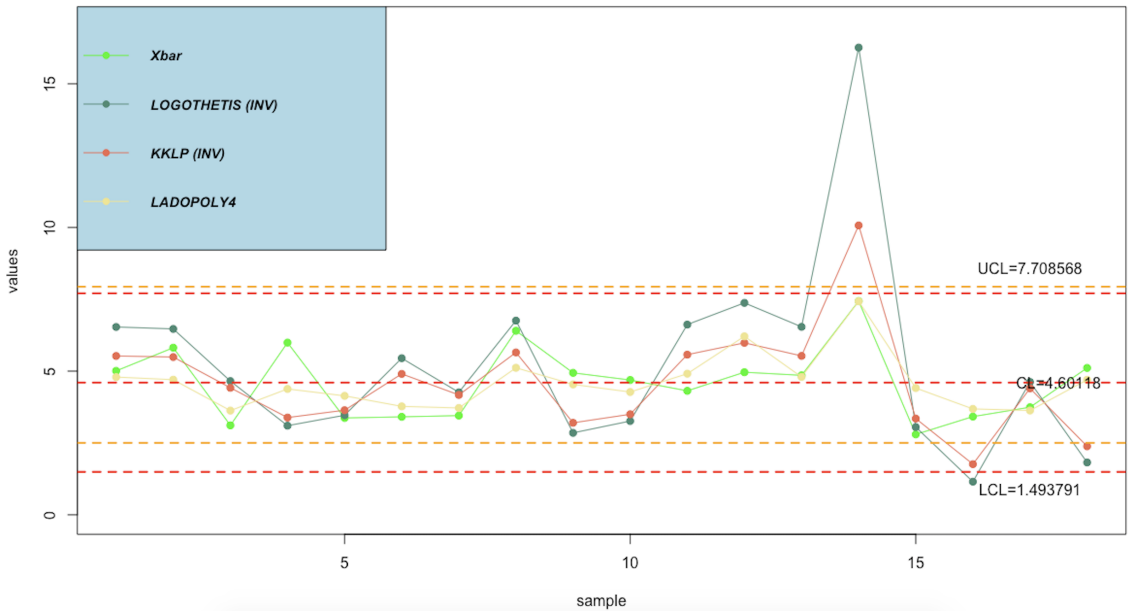

Figure 4.1: $\overline{X}$-Graph

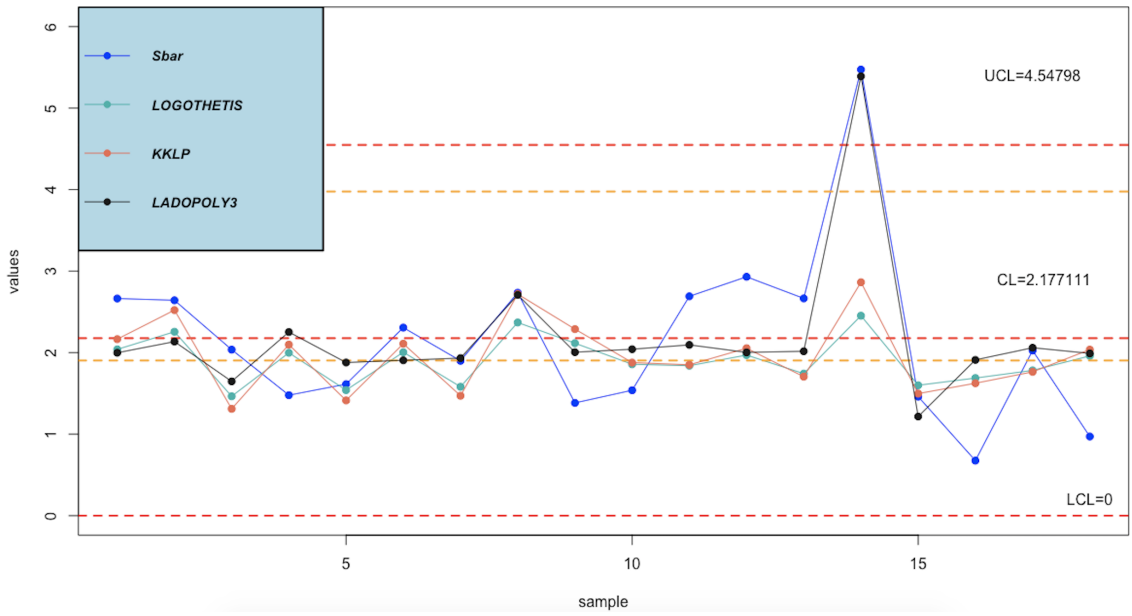

Figure 4.2: S-Graph

Note that "Ladopoly4" and "Ladopoly3" in Figures 4.1 and 4.2 describe the degree k of the polynomial presented in (4.6) i.e., 4 and 3, respectively.

4.4 Conclusions

Data transformation constitutes a crucial component of statistical quality control since it helps in achieving a satisfactory degree of homoskedasticity and at the same time ensuring high accuracy and great applicability.
In this chapter, we suggested two adjusted transformations:

1. The **KKLP** method, which is described by the non-linear model in equa-

tion (4.1), manages to identify and portray variations of noise behaviours through the $\overline{X}-$Graph under the transformation

$$\mu = \left(\frac{\sigma}{\alpha}\right)^{1/k}, \tag{4.7}$$

and hence, the double $\overline{X}$ control chart ($\overline{X}$, KKLP**(INV)**) is designated which is enough to display the patterns of both noise and mean of the data. The same evaluation holds for the double S-chart (S, KKLP). Also, the KKLP succeeds in resolving entirely confusion issues in ANOVA.

2. The **Ladopoly** method, which is described by the linear model in equation (4.6), attempts to describe as accurately as possible the original $\overline{X}$, S charts. Also this method succeeds in identifying OOC points and being smooth enough for the IC points.

Bibliography

[1] Abebe, A. (2000). Introduction to design and analysis of experiments with the SAS system. *Discrete and Statistical Sciences, Auburn University.*

[2] Antzoulakos, D. (2006). Statistical quality control. *University notes, Post graduate program "Applied Statistics"*, *University of Piraeus*

[3] Bickel, P.J. and Kjell, A.D. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, **76(374)**, 296-311.

[4] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **26(2)**, 211-252.

[5] Chao, M.T. and Cheng, S.W. (2008). On 2-D control charts. *Quality Technology and Quantitative Management*, **5(3)**, 243-261.

[6] Chao, M.T. and Smiley, W.C. (1996). Semicircle control chart for variables data. *Quality Engineering*, **8(3)**, 441-446.

[7] Chen, G. and Smiley, W.C. (1998). Max chart: combining X-bar chart and S chart. *Statistica Sinica*, **8**, 263-271.

[8] Cook, R.D. and Weisberg, S. (1999). Applied regression including computing and graphics. Wiley, New York.

[9] Davis, C.S. (2002). Statistical methods for the analysis of repeated measurements. *Springer Science & Business Media.*

[10] Duncan, A.J. (1959). Quality control and industrial statistics (Rev. Ed.). *Richard D. Irwin, Inc.*, Homewood, Illinois, 161-170.

[11] Fisher, R. A. (1919). XV.−The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, **52(2)**, 399-433.

[12] Fox, J. and Weisberg, S. (2019). An R companion to applied regression. *Sage*, UK.

[13] Gary, W.O. (2010). A first course in design and analysis of experiments. *WH Freeman*, New York, USA.

[14] Girden, E.R. (1992). ANOVA: Repeated measures. No. 84. *Sage publications*, Thousand Oaks.

[15] Haim, S. (2005). Response modeling methodology: empirical modeling for engineering and science. Vol. 8. *World Scientific*.

[16] Hawkins, D.M. and Weisberg, S. (2017). Combining the Box-Cox power and generalized log transformations to accommodate non-positive responses in linear and mixed-effects linear models. *South African Statistical Journal*, **51(2)**, 317-328.

[17] John, J.A. and Norman, R.D. (1980). An alternative family of transformations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **29(2)**, 190-197.

[18] Juran, J. and Godfrey, A.B. (1999). Quality handbook. *Republished McGraw-Hill*, New York.

[19] Kackar, R.N. (1985). Off-line quality control, parameter design, and the Taguchi method. *Journal of Quality Technology*, **17(4)**, 176-188.

[20] Lane, D. (2008). Online statistics: An interactive multimedia course of study.

[21] Logothetis, N. (1990). Box–Cox transformations and the Taguchi method. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **39(1)**, 31-48.

[22] Logothetis, N. (2002). Continuous Quality Improvement Procedures and Techniques. *(Hellenic open University, Patra)*.

[23] Manly, B.F.J. (1976). Exponential data transformations. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **25(1)**, 37-42.

[24] McCracken, A.K. and Chakraborti, S. (2013). Control charts for joint monitoring of mean and variance: an overview. *Quality Technology & Quantitative Management*, **10(1)**, 17-36.

[25] Montgomery, D.C. (2009). Introduction to statistical quality control. Vol 7. *John Wiley & Sons*.

[26] Montgomery, D.C. (2017). Design and analysis of experiments. *John Wiley & sons*.

[27] Nesselroade, K.P. and Grimm, L.G. (2019). Statistical applications for the behavioral and social sciences. *Wiley*.

[28] Pengfei, L. (2005). Box-Cox Transformations: An overview. *Documento de trabajo, EEUU, Department of Statistics, University of Connecticut*.

[29] Quinn, G.P. and Keough, M.J. (2002). Experimental design and data analysis for biologists. *Cambridge University Press*.

[30] Rakitzis, A. (2016). Statistical quality control. *Lecture Notes, University of the Aegean*.

[31] Ryan, T.P. (2011). Statistical methods for quality improvement. *John Wiley & Sons*.

[32] Sakia, R.M. (1992). The Box-Cox transformation technique: a review. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **41(2)**, 169-178.

[33] Shewhart, W.A. (1931). Economic control of quality of manufactured product. *Martino Fine Books*, Eastford.

[34] Taguchi, G. and Konishi, S. (1987). Taguchi methods orthogonal arrays and linear graphs: Tools for quality engineering. *American Supplier Institute*, Egypt.

[35] Tsui, K. (2002). An overview of Taguchi method and newly developed statistical methods for robust design. *lie Transactions*, **24(5)**, 44-57.

[36] Tukey, J.W. (1957). On the comparative anatomy of transformations. *The Annals of Mathematical Statistics*, **28(3)**, 602-632.

[37] Yeo, I.K. and Johnson, R.A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, **87(4)**, 954-959.