A STRATEGY IN THE STATISTICAL ANALYSIS OF THE EXPERIMENTAL DESIGN

Nidia Hernández Pérez, Vivian Sistachs Vega y Elina Miret Barroso
Faculty of Mathematics and Computation, University of Havana
Miguel Angel Díaz Martínez, Department of General Mathematic, ISPJAE

ABSTRACT
Additivity, variance, homogeneity and normality of the errors are often violated in the ANOVA of experimental design. Considering the results in the literature, we propose a general strategy in statistical analysis of an experimental design. This strategy may select the adequate transformation and model when the basic model assumptions are violate.

Key words: model selection, ANOVA, homogeneity Tests.

RESUMEN
Frecuentemente las hipótesis sobre aditividad, homogeneidad de varianzas y normalidad son violadas en el ANOVA de los diseños experimentales. Considerando los resultados de la literatura poponemos una estrategia general en el análisis estadístico de un diseño experimental. Esta estrategia puede seleccionar la transformación adecuada y modelar cuando las asunciones básicas son violadas.

Palabras clave: selección ejemplar, ANOVA, pruebas de homogeneidad.

INTRODUCTION

In order to include ANOVA to statistical analysis applied to the results from an experimental design, it is necessary to check some requirements as normality, independence, constant variance and additivity of the effects and interactions of the controlled variables in the designed experiment.

M.S. Barttlet (1947), F.I. Anscombe, S.W. Tuckey (1963), G.E.P. Box, D.R. Cox (1964), W.J. Canover, R. Duncan (1981), G.E.P. Box, W.G. Hunter and Hunter (1987), G.C.I. Fernández (1992) and other, have suggested several procedures to give solution to problems where ANOVA assumptions are not satisfied. The more used procedures are:

1. Data transformations to power by Box and Cox procedure (1967),
2. Transformations to range suggested by Canover and Duncan (1981),
3. Non parametric methods.

When the violations are moderately satisfy, the solution 1 and 2 are applied. However, the transformed data may or not satisfy the ANOVA assumptions. It these are satisfy, we can consider models with the new data and make the validation analysis for each model and choice it according to some criterion. If the violations are very strong, the literature suggests the solution 3.

Frequently ANOVA procedures are specially robust when the number of repetitions is the same and the violations are not strong.

General strategy

Applying the general strategy considered by C. Chalfield (1988), we propose the combination of the following phases:

(I) Preliminary and exploratory analysis.

(II) Conformatory analysis.

(III) Selection of models.

The phase (I) pretends to explore the data behavior and to establish relationship among the levels of the controlled variables in the experiments without inferential methods. This work has to part, the first consists in the study of the structure of data, measure scale, catch on the computer, and the second includes graphics of the treatment means, boxplot to treatments in order to see and compare the behavior of the outcome variable under the experimental conditions. Normality and homogeneity of data can be detected by another graphics as normal plot. In this part is necessary the detailed observation of these results by the research permitting to make prior conclusions before to apply confirmatory test and the selection of models to fit.

With this initial analysis, we can stablish that:

1. If there exist violations or not in the ANOVA assumptions.

2. Determinations of possible outliers and observation errors.

The phase (II) consists in the applications of a set of classic statistical tests that will permit the confirmation of the conclusions of the phase (I). We should remember that all these tests take place in a probabilistics model and they can be sensitive to its violations.

When one makes the decision of the transformation it is necessary to repeat these two phases before to fit models to phase (III). The behavior of residues for each model in the phase (III) is fundamental to infer if there are or not some violations after to fit models.

It is important in this strategy to consider a careful analysis of original data and transformed data and the correspondent residual analysis in the fitted models.

In the practice, we can have a transformations set given 1) and 2) where some or all the ANOVA assumption do not satisfy. These transformations give a models set and the problem is what combination of transformation and model is better in some sense.

CRITERION OF MODEL SELECTION WITH TRANSFORMATIONS

Let be: T a set of transformations t, $M_t(x)$ the model set of the corresponding to the transformation $t \in T$ for the controlled variable, that is:

$m_t(x) \in M_t(x)$ if $t(y) = m_t(x) + e$, with $t \in T$ and $e \sim N(0,\sigma^2(t))$ independent and $t \in T$.

Suposse that $T' \subset T$ and $M_{T'}(x) \subset M_t(x)$ such that the elements of $M_{T'}(x)$ satisfy the assumptions of the ANOVA. Then, a simple criterion of selection of t' and $m_{t'}(x)$ can be given as follow:

$$m^*_{t'}(x) = \arg \min_{\substack{t' \in T' \\ m_{t'}(x) \in M_{T'}(x)}} SSR(m_{t'}(x))$$

with $SSR(m_t(x))$ is the Residual Sum the Squares correspondent to $m_t(x)$

REAL DATA EXAMPLE

This is a work about the sensibility of AROMA artificial pasture to different pesticides.

AROMA artificial pasture was studied by a team from the PASTURE DEPARTMENT of the Animal Science Institute, Havana, Cuba.

To obtain the data, they took four blocks to apply four kinds of pesticides. Each block was divided in four sections, making a total of twenty experimental units.

In each experimental unit the number of born plants, dead plants and plants that were kept alive at the end, were measured.

A block was considered a homogeneous unit.

DATA TABLE (with the number of plants)

	B O R N				**D E A D**				**ALIVE AT THE END**			
TREATMENT	**I**	**I I**	**I I I**	**IV**	**I**	**I I**	**I I I**	**IV**	**I**	**I I**	**I I I**	**IV**
Bioester	14	14	15	22	0	1	2	2	14	13	13	20
Butephan	16	13	18	10	1	0	2	0	15	13	16	10
Atrazina	22	18	16	24	22	18	16	24	0	0	0	0
Duiran	14	14	17	19	2	0	5	4	12	14	12	15
Control	14	14	13	19	0	0	0	0	14	14	13	19

An important results analysis inside the exploratory analysis we the principals statistic for dead plant (ANEX 1). Here we look the were variability between to data in fact the tested for homogeneity of variances is not significative, beside in this exploratory study we observe homogeneity with block and the diference with the treatment. It should be remarked that our outcome variable has a Poisson distribution because it is a count variable. This indicates that the assumptions of the model:

$$Y_{ij} = \mu + \alpha_j + \beta_j + e_{ij}, \qquad e_{ij} \sim N(0,\sigma^2) \qquad V(e_{ij}) = \mu + \alpha_j + \beta_j \ \forall_{ij}$$

are violated, especially the normality and homogeneity and was reflected in the normal plot (ANEX 3). In ANEX 1 it is possible to see the non-homogeneity of variances. The Literature recommends in this case to make the following transformation: $y = \sqrt{x + 0.375}$. The descriptive analysis with the transform data is resumed in (ANEX 2).

ANEX 4 shows the box-plot results to treatments and to blocks. We remark the homogeneity between blocks and between treatments except the atrazina whose bahavior is substantially different.

After the transformation it was obtained homogeneity of variances and normality. Thus, a new problem is to find if there exist some effects of the treatment or not.

The results in this part of the work tolds us that only the dead plants had some effects of the treatment and that the behavior of the ATRAZINA was the worst of them. All above, permitted us to conclude that the sensibility of AROMA pasture to atrazina was the most of all pesticides used in the experiment.

ANEX 1

Variable	SAROMAMU.y
Sample size	16
Average	6.1875
Median	2
Mode	0
Variance	72.4292
Standard deviation	8.51053
Minimum	0
Maximum	24

Analysis of variance for SAROMAMU.y

Source of variation	Sum of Squares	df	Mean square	F-ratio	Sig. level
Main Effects	1041.3750	6	173.56250	34.664	0.0000
SAROMAMU. tratamient	1026.1875	3	342.06250	68.318	0.0000
SAROMAMU. bloque	15.1875	3	5.06250	1.011	0.4319
Residual	45.062500	9	5.0069444		
TOTAL (corr.)	1086.4375	15			

0 missing values have been excluded.

Test for Homogeneity of Variances

Cochran´s test: 0.6639 p = 0.0651417
Bartlett´s test: 1.84297 p = 0.0919851
Hartley´s test: 14.5455

ANEX 2

Variable	SAROMAMU.ytransfor
Sample size	16
Average	2.08571
Median	1.54113
Mode	0.61238
Variance	2.3604
Standard deviation	1.533636
Minimum	0.61238
Maximum	4.93711

Analysis of variance for SAROMAMU.ytransfor

Source of variation	Sum of Squares	df	Mean square	F-ratio	Sig. level
Main Effects	33.170816	6	5.528469	22.261	0.0001
SAROMAMU. tratamient	31.981510	3	10.660503	42.926	0.0000
SAROMAMU. bloque	1.189306	3	0.396435	1.596	0.2576
Residual	2.2351235	9	0.2483471		
TOTAL (corr.)	35.405940	15			

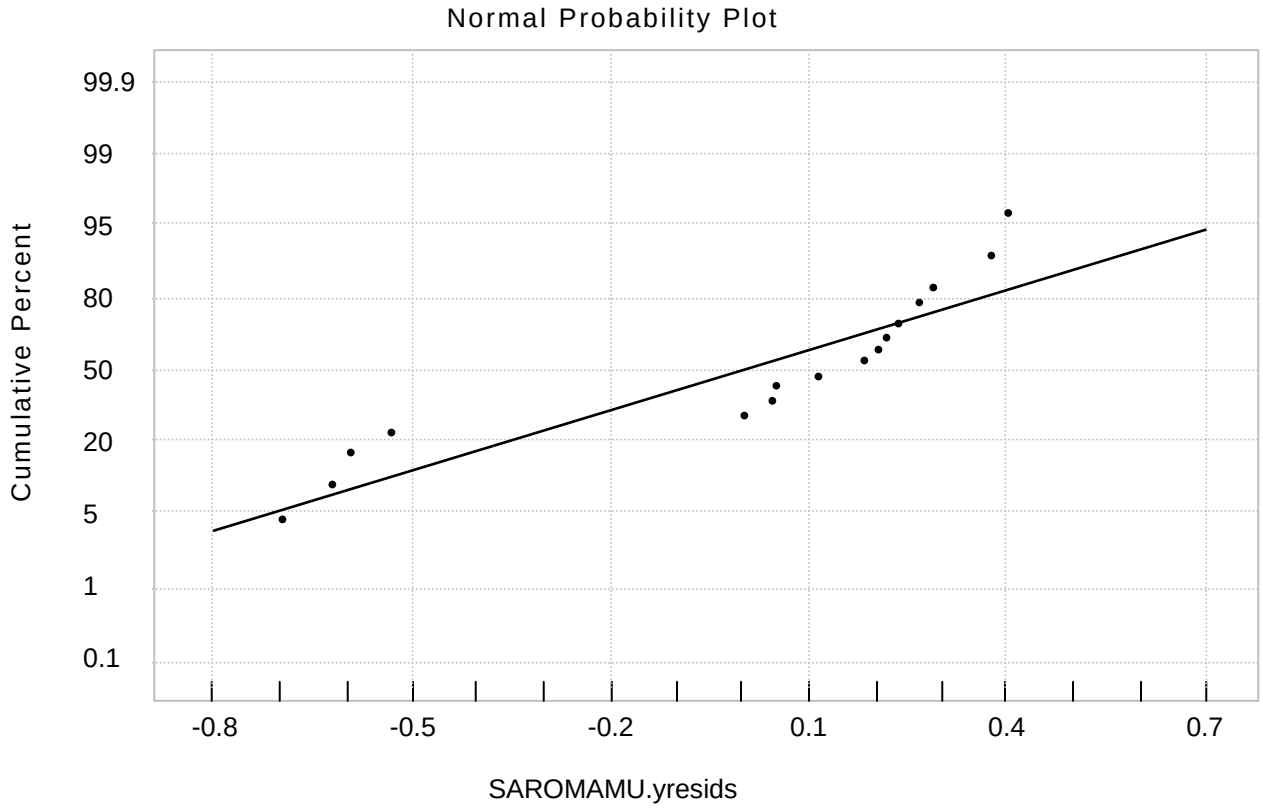
0 missing values have been excluded.

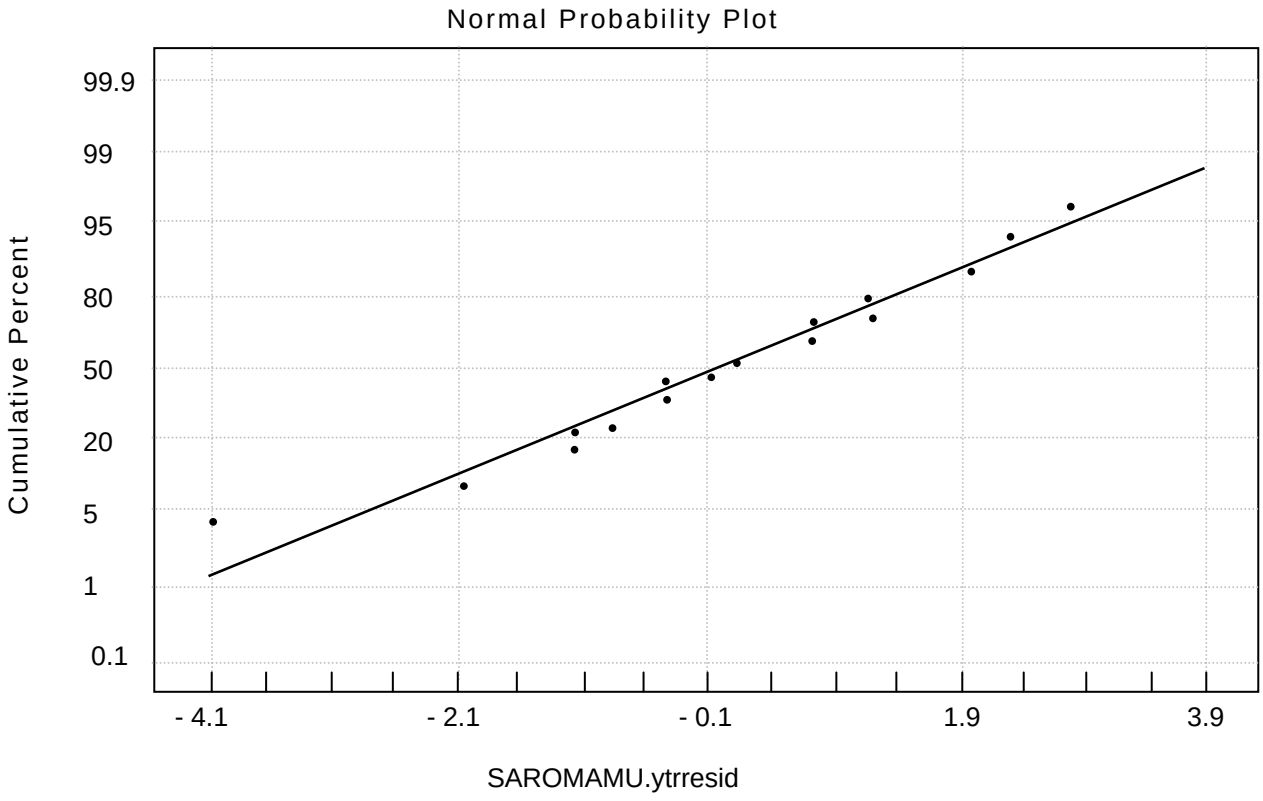
Test for Homogeneity of Variances

Cochran´s test: 0.505021 p = 0.336459
Bartlett´s test: 1.14939 p = 0.689907

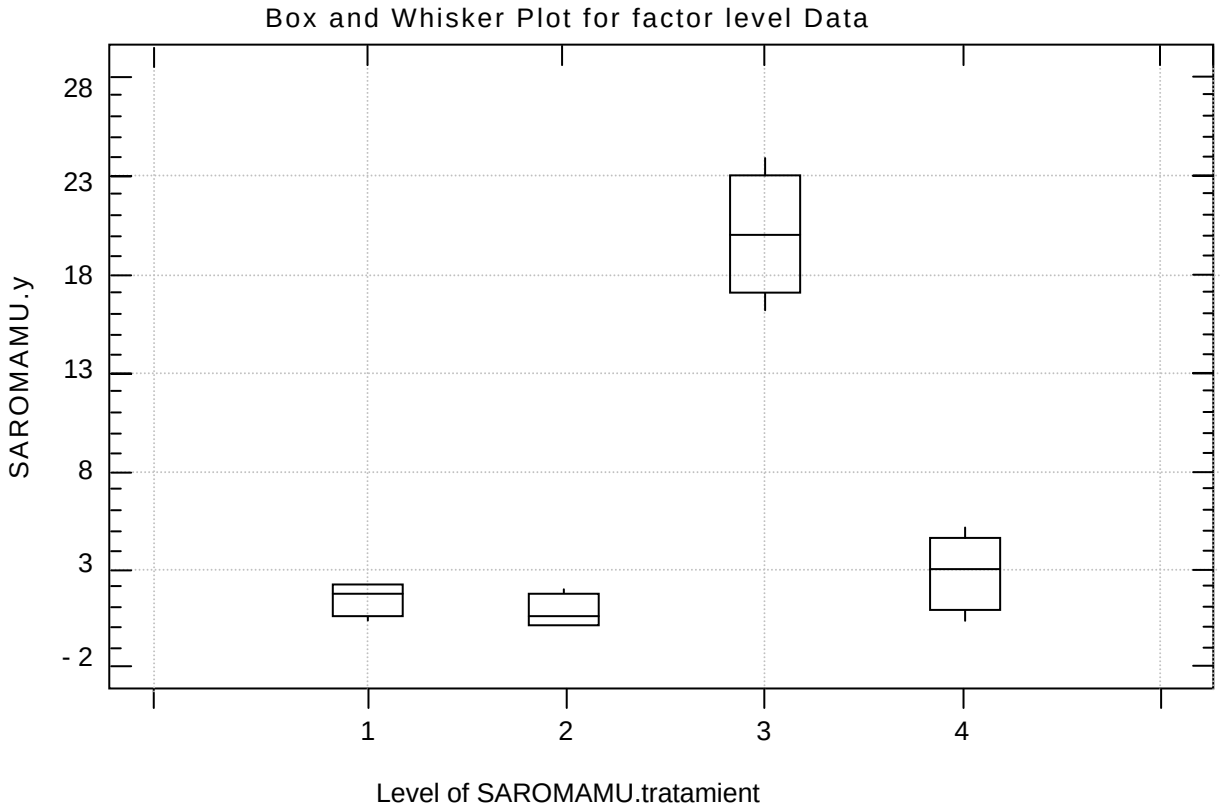
Hartley´s test: 3.49253

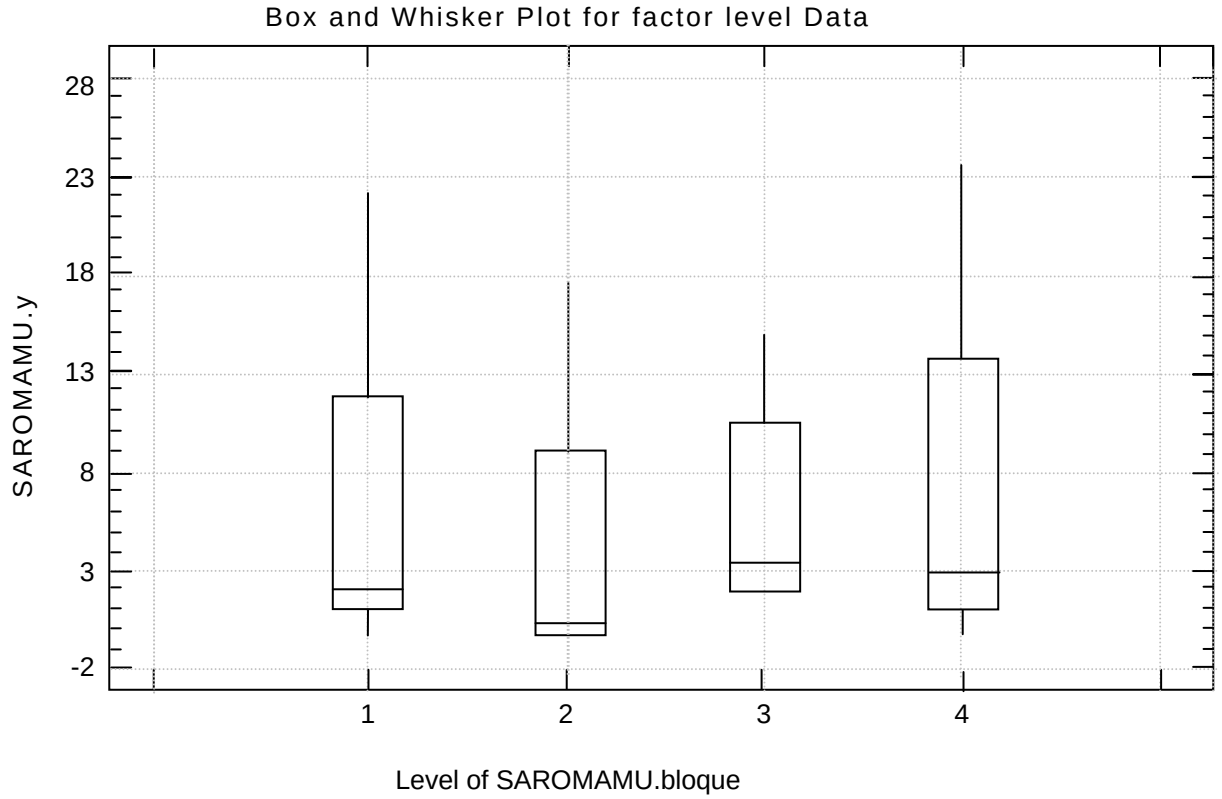
ANEX 3

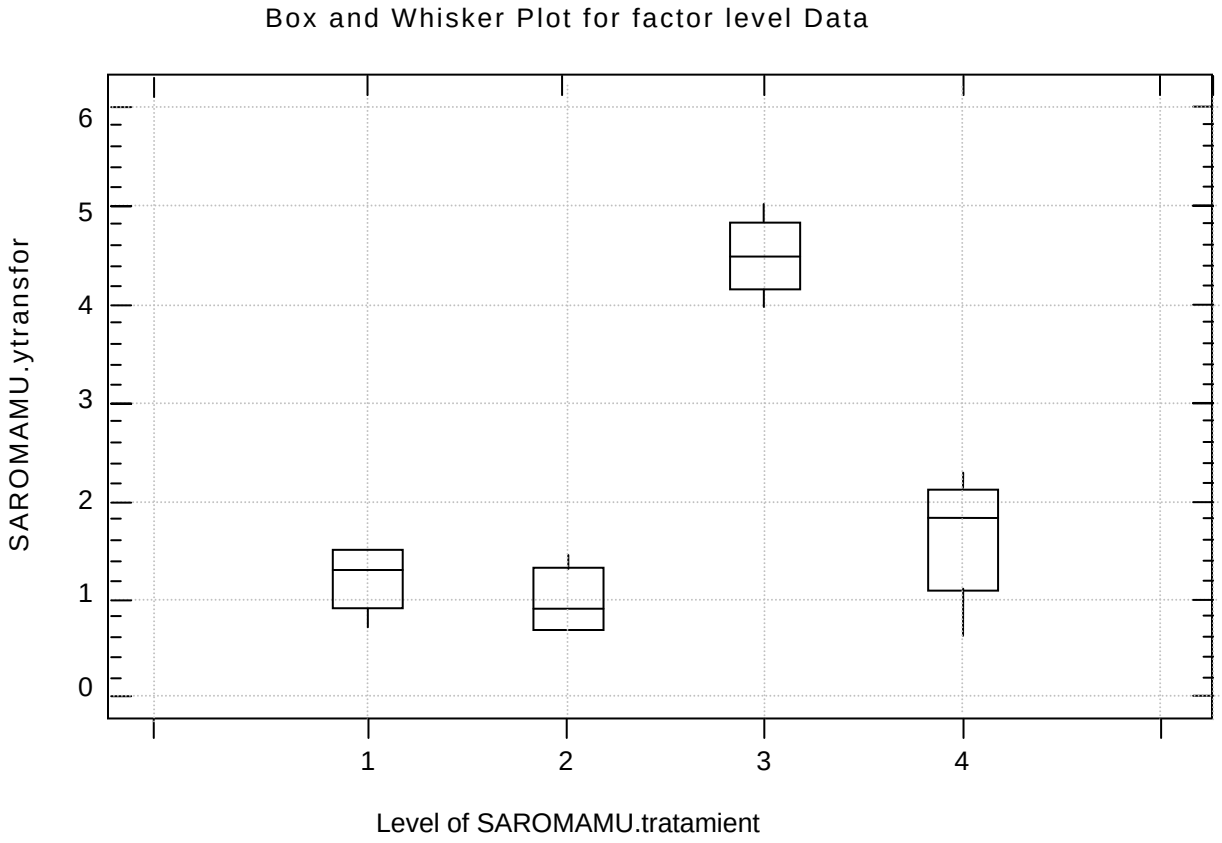


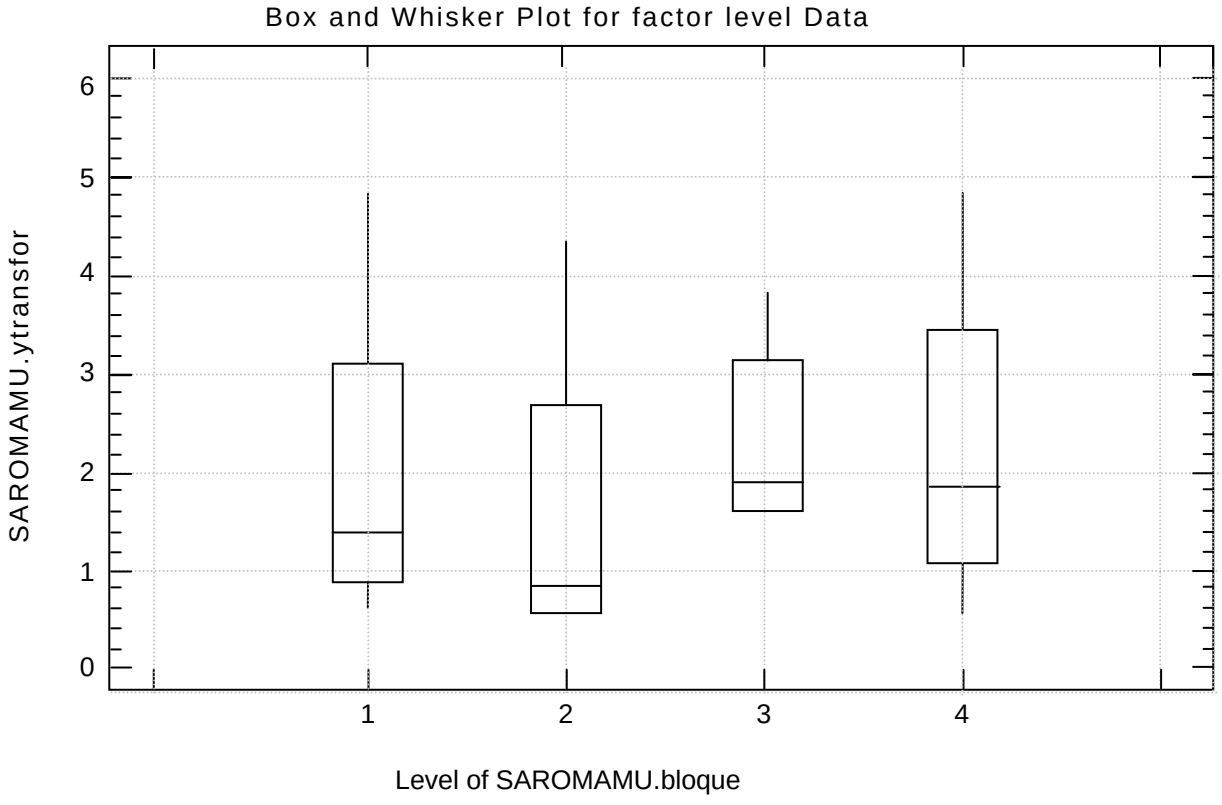


ANEX 4


Box and Whisker Plot for factor level Data
SAROMAMU.y
28
23
18
13
8
3
- 2
1
2
3
4
Level of SAROMAMU.tratamient


Box and Whisker Plot for factor level Data
SAROMAMU.y
28
23
18
13
8
3
-2
1
2
3
4
Level of SAROMAMU.bloque


Box and Whisker Plot for factor level Data
SAROMAMU.ytransfor
0
1
2
3
4
5
6
1
2
3
4
Level of SAROMAMU.tratamient


Box and Whisker Plot for factor level Data
SAROMAMU.ytransfor
0
1
2
3
4
5
6
1
2
3
4
Level of SAROMAMU.bloque

REFERENCES

ANSCOMBE, F.I. and W. TUCKEY (1963): "The examination and analysis of residual", **Technometrics** 5-XI-160.

BARTLETT, M.S. (1947): "The use of transformations", **Biometrics** 3, 39-52.

BOX, G.E.P. and D.R. COX (1964): "An analysis of transformation". JRSS Seri B 26, 221-243.

_____; W.G. HUNTER and M.T.S. HUNTER (1987): "Estadística para investigadores. Una introducción al Diseño, Análisis de datos y construcción de modelos". Editorial REVERTE.

CANOVER, W.S. and E.L. IMAN (1991): "Rank transformation as a bridge between parametric and non parametric statistics". **A. Statistician** 35, 124-129.

CHATFIELD, C. (1988): "Problem Solvin". **A. Statistician Guide**. Chapman and Hiel.

FERNANDEZ G., C.J. (1992): "Residual and Data transformation. Important tools in Statistical Analysis". **Howrt Science** 4, April, 292-309.