

Some combinatorial structures in experimental design: overview, statistical models and applications

Abstract

Background: Design and analysis of experiments will become much more prevalent simultaneously in scientific, academic and applied aspects over the next few years. Combinatorial designs are touted as the most important structures in this field taking into account their desirable features from statistical perspective.^{1,2} The applicability of such designs is widely spread in areas such as biostatistics, biometry, medicine, information technologies and many others. Usually, the most significant and vital objective of the experimenter is to maximize the profit and respectively to minimize the expenses and moreover the timing under which the experiment take place. This necessity emphasizes the importance of the more efficient mathematical and statistical methods in order to improve the quality of the analysis.

We review combinatorial structures,³ in particular balanced incomplete block design (BIBD)⁴⁻⁶ and Latin squares designs (LSD),⁷⁻⁹ which were first introduced by R.A Fisher and et al. in 1925, who developed the basic statistical theory of such designs.

We propose general framework, using the mathematical structures in Experimental design, to demonstrate those combinatorial designs which sometimes can be easily constructed by dint of computer tools.¹⁰ Applications on Biostatistics and Biometry fields are illustrated, namely an example dealing with the comparison of pharmacological substances in terms of reaction time in a bio-statistical experiment and another one dealing with comparisons of clinical effects of a new medical product. Simulations and statistical analysis are presented using R Studio and the variety of built packages related to Design of Experiment.^{11,12}

Keywords: balanced incomplete block design, design of experiments, latin square, r statistics, biostatistics, biometry

Volume 7 Issue 4 - 2018

Petya Valcheva,¹ Teresa A Oliveira²

¹Department of Probability, Sofia University, Bulgaria

²Departamento de Ciências e Tecnologia, Universidade Aberta, Portugal

Correspondence: Petya Valcheva, Department of Probability, Operations research and Statistics, Sofia University "St. Kliment Ohridski", Student's Town building 55, entrance V, Bulgaria, Tel +3598 9665 4485, Email pvalcheva@unwe.bg

Received: July 02, 2018 | **Published:** August 10, 2018

Introduction

Design of experiments (DOE) is an important branch of applied statistics that deals with planning, conducting of the experiment, analyzing and interpreting final results. It combines mathematical and statistical tools, which aim at constructing optimal designs to be tested. Due to the widely application during recent decades, this science is strongly spread in many areas such as optimization, process quality control as well as product performance prediction.

The historical notes highlight that some of the most remarkable and progressive contributions of statistics in the twentieth century have been those in Experimental design. British statistician and geneticist Sir R.A. Fisher first laid the foundations in this area, between 1918 and 1940, as a result of different applications and simulations in agricultural experiments. Most of his early publications have emphasized the fact that profound conclusions could be drawn efficiently from fluctuations of nuisance variable such as fertilizers, temperature and other natural conditions. Similar methods have been successfully applied to variety of areas in order to investigate the effects of many different factors by changing them at one time instead of changing one factor at a time.

Next significant period, also known as "The First Industrial Era", occurred as a result of the application of experimental designs in chemical industry. It was elaborated in the 1950s till late nineties by the extensive work of G.Box and B.Wilson on the famous *Response Surface Methodology (RSM)* that explores the relationships between

several explanatory variables and one or more response variables. Over the past years there has been a tremendous increase in the exploitation similar experimental techniques in optimization processes and industry. This is due largely to the increased emphasis on quality improvement and the essential role played by statistical methods used in DOE.

"The Second Industrial Era" was conceived in late 1970s after the exhaustive work of the Japanese quality consultant Genichi Taguchi. His *Robust Design method (RDM)* was the leading approach in quality improvement methods focused on response surfaces associated with both mean and variation reduction, and to choose the noise factor settings, so that both variability and bias are made simultaneously small.

Experimental design techniques are effective and powerful methods that are also becoming popular in the area of computer-aided design and engineering using computer simulation models. Some basic properties as maximizing the amount of information while minimizing the amount of the collected data have had revolutionary impact among scientists. This fact allows us to lay the foundations of the "Modern Era", beginning circa 1990, when the design techniques have been also becoming popular in different sectors of economy.

We perform description of two combinatorial structures, namely Balanced incomplete block design and Latin square design, and demonstrate its application in statistical analysis. Practical methods for analyzing data from life testing will be provided for each design.

We focus on planning experiments efficiently and how to create statistical analysis with the aid of R packages for experimental design.

Generating balanced incomplete block designs

Block designs arise in experimental design as fundamental units for testing too many varieties in an experiment. Such constructions can efficiently provide information in cases when treatments are included in blocks, because they are expensive or the testing time should be minimized. However, sometimes blocks or experiment's budget may not be large enough to allow all desirable treatments to be executed in all blocks. The incomplete block designs refer to the condition when each block has less than a full complement of treatments. But the most intensely studied are the balanced incomplete block designs (BIBDs or 2-designs), in which all treatment effects and their differences are estimated with the same precision as long as every pair of treatments occurs together the same number of times. The statistical analysis of such designs is considerably more complicated, although they are used in cases having one source of variation.

A *Balanced incomplete block design* with parameters (v, b, r, k, λ) is an ordered pair (V, B) , where V is a finite v -element set of treatments or varieties, B is a family of k -element subset of V , called blocks such that satisfy the following conditions:

- (i) Each block contains exactly k members
- (ii) Every treatment is contained in exactly r blocks (or is replicated r times)
- (iii) Every 2-subset of V (pair of treatments) is contained in the same block exactly λ times.

$BIBD(v, b, r, k, \lambda)$ is an arrangement of b subsets of size k from a set of v treatments, such that (i), (ii) and (iii) are satisfied. The parameter λ must be an integer. The necessary, but not sufficient conditions for the existence of a *BIBD* are:

$$vr = bk \lambda (v - 1) = r(k - 1)b \geq v$$

If $v = b$, the *BIBD* is said to be a symmetric.

There are different packages in R for creating and analyzing experimental designs for research purposes. The package "crossdes" generate cross-over designs of various types, including Latin squares and BIBD. The build-in function "find.BIB" gives rise to design with desired parameters, where number of rows corresponds to the blocks and columns - the number of elements per block. The R output gives the following result: (Figure 1).

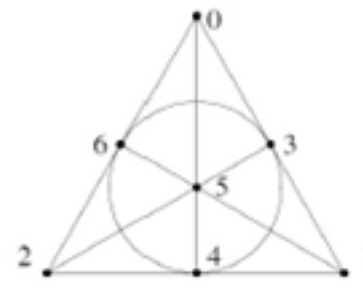
The resulting design can be verified concerning balanced manner via "isGYD" function. The conclusion is shown below:

- i. isGYD (find.BIB(7,7,3))
- ii. The design is a balanced incomplete block design w.r.t. rows.

There are also other packages in R, which can be used for generating block designs. For example "ibd", "AlgDesign" and "dae".

The above design, $BIBD(7, 7, 3, 3, 1)$, which is symmetric ($v = b = 7$), corresponds to the *Steiner triple system* of order 7 ($STS(7)$) consists of a set V of 7 points, and a collection B of subsets of V called triples, such that each block contains exactly 3 points, and any two points lie

together in exactly one block. This system has cyclic representation: let the set $V = \{0, 1, \dots, 6\}$ be the integers mod 7 and the triples are the set $\{1, 2, 4\}$ of quadratic residues mod 7 and its cyclic shifts. The system is also known as the projective plane of order 2, or the *Fano plane*, which has the smallest possible number of points and lines - 3 points on every line and 3 lines through every point.



>find.BIB(7,7,3)

	[,1]	[,2]	[,3]
[1,]	1	2	4
[2,]	2	3	5
[3,]	3	4	6
[4,]	4	5	7
[5,]	5	6	1
[6,]	6	7	2
[7,]	7	1	3

Figure 1 Fano plane.

Statistical analysis of numerical example

Balanced incomplete block designs are typically used when all comparisons are equally important for the experiment, but the researcher is not able to run all possible combinations. In such cases the treatments that are used in each block should be selected in *balanced manner*, i.e. any pair occurs together in the same number of times as any other pair.⁶

Consider a $BIBD(v, b, r, k, \lambda)$ that satisfies conditions (i), (ii) and (iii).

The statistical model of the design is

$$y_{ij} = \mu + \tau_i + \beta_j + \xi_{ij} \quad i = 1, \dots, v \quad j = 1, \dots, b$$

where

y_{ij} is the i^{th} observation in the j^{th} block

μ is the general mean effect

τ_i is the effect of the i^{th} treatment

β_j is the effect of the j^{th} block

ξ_{ij} the i.i.d. random error component with $\xi_{ij} \sim NID(0, \sigma^2)$

In the following illustrative example, we will use already generated design $BIBD(7, 7, 3, 3, 1)$. Suppose an experiment is to be run to compare $v = 7$ compositions of pharmacological substances in terms

of reaction time in the bio statistical experiment. Further, assume that only 3 observations can be taken per day, and that the experiment must be completed within 7 days. The incidence matrix in Table 1 below has a 1 in the (i, j) -cell, if the treatment i is contained in block j and 0 otherwise, and also the data of above-mentioned example (Table 1).

For this experiment we apply Inter-block analysis of variance, where the treatment effects are estimated after eliminating the block effects from the normal equations. When blocks are incomplete, there are two sources of information about treatment effects, but the bigger part comes from the analysis done below. In Table 2, we give variance table about such analysis that can be compiled into the intra - block analysis of variance table for testing the significance of treatment effect given as follows: (Table 2).

The form of the ANOVA used to analyze BIBD data depends on the type of analysis. After its application, the researcher retains or rejects the hypothesis, often based on a statistical mechanism called hypothesis testing. The null hypothesis of our interest is

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_7$$

and the alternative hypothesis is

$$H_1 : \text{at least one pair of } \tau_j \text{'s is different}$$

In Table 3 we present output results, estimated using some basic functions in R. For example, the linear model function $lm()$ to conduct linear regression analysis and $anova()$ function as a traditional statistical approach (Table 3).

Table 1 Incidence matrix for BIBD (7,7,3,3,1)

Blocks							
Treatments	1	2	3	4	5	6	7
1	1(=73)	0	0	0	1(=64)	0	1(=66)
2	1(=71)	1(=68)	0	0	0	1(=65)	0
3	0	1(=67)	1(=72)	0	0	0	1(=72)
4	1(=75)	0	1(=74)	1(=73)	0	0	0
5	0	1(=71)	0	1(=69)	1(=70)	0	0
6	0	0	1(=68)	0	1(=67)	1(=71)	0
7	0	0	0	1(=71)	0	1(=75)	1(=74)

Table 2 Intra-block analysis of variance table for BIBD

Source of variation	Sum of squares	Degrees of freedom	Mean square	F ₀
Between Treatments				
(adjusted)	SS _{Tr(adj)}	v-1	MS _{Tr} =SS _{Tr(adj)} /v-1	MS _{Treatment} /MS _E
Between Blocks				
(adjusted)	SS _{Blocks(unadj)}	b-1		
Intrablock				
Error	SS _{Error(substraction)}	N-a-b+1	MS _{Error}	
Total	SS _{Total} = $\sum \sum y_{ij}^2 - G^2/N$	N-1		

Table 3 Analysis of variance table

	Df	Sum of squares	Mean square	F value	Pr(>F)
Blocks	6	85.619	14.2698	1.1515	0.4146
Treatments	6	29.524	4.9206	0.3971	0.8617
Residuals	8	99.143	12.3929		

The test for null hypothesis H_0 is based on the rule that if $F_{TR} > F_{1-\alpha, v-1, bk-b-v+1}$, then H_0 is rejected. At the 5% significance level, the p -value for treatments is less than 0.05, which means that the null hypothesis is rejected and the difference between group considering the blocks is not significant as well (p -value > 0.05).

Regression analysis is a very powerful tool for better understanding the relationship between one or more predictor variables and the response variable. When we run such model, the variance of the errors must be constant and they must have a mean of zero. If this isn't the case, the model may not be valid. To verify these assumptions, we

should check the model adequacy that includes the verification of the independence and normality of the residuals. Below are the plots from the analysis we do for the numerical example: The first graph illustrates residuals versus fitted values from the standard regression model for BIBD. The errors have constant variance, with the residuals scattered randomly around 0. If the residuals increase or decrease with the fitted values, the errors may not have the constant variance. The third Normal Q-Q plot (Quantile - Quantile plot) indicates the normality of the residuals $e_{ij} = y_{ij} - \tilde{y}_{ij}$. The second picture shows residuals in case when variance is more constant. We emphasize that the regression model is transformed concerning a logarithmic function (Figure 2).

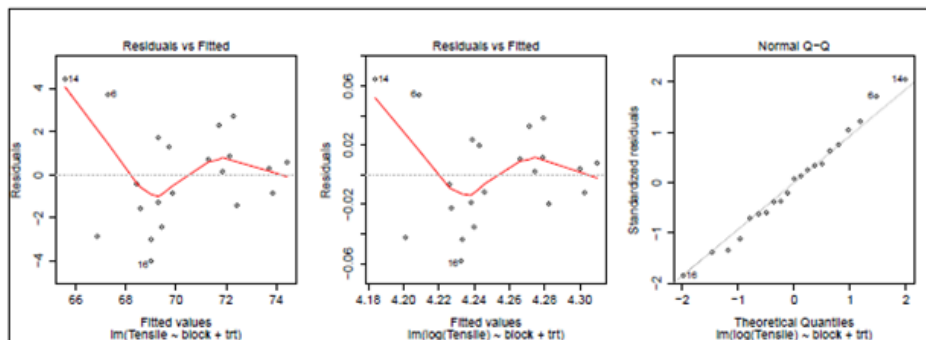


Figure 2 Checking assumptions about residuals in regression analysis.

Statistical computing for Latin square design

In this section we consider the brief history of Latin square designs (LSDs), the basic statistical model and analysis of variance table, and finally a numerical example, estimated using R and appropriate packages. In 1782, the famous Swiss mathematician Leonhard Euler first introduced Latin squares in his famous entertaining *Thirty-six Officers problem*: Given 6 distinct regiments each consisting of 6 distinct ranks, is it possible to arrange a (6×6) grid such that each row and each column of the grid contains exactly one representative from each regiment and exactly one representative of each rank? After so many years, this problem is still unsolved and is conjectured that there was no such arrangement. But on the other hand, it is believed that the question marks the beginning of the progressive investigation of Latin squares.^{10,12}

A Latin square of order n is an $n \times n$ array consisting of n distinct symbols from a set N of cardinality n , such that each symbol appears exactly once in each row and exactly once in each column. Such efficient designs are primarily used in Experimental design, in particular in agricultural, biological and medical experiments. The use of LSD seems to be highly effective for controlling two source of external variation. The principle can be further extended to control more than two sources of variation. The design is also useful for investigating simultaneously effects a single treatment and two possible blocking variables, each with the same number of levels.

The statistical model for a Latin square is

$$y_{ijk} = \mu + \alpha_i + \tau_j + \beta_k + \xi_{ijk}, i, j, k = 1, \dots, p$$

where:

y^{ijk} is the observation in the i th row and k th column for the j th treatment

μ is the overall mean

α^i is the i th row effect

τ^j is the j th treatment effect

β_k is the k th column effect

ξ_{ij} is the random error component

Consider an experiment conducted to investigate the clinical effect of a new medical product. Four volunteers were given varying doses from the medicine and each of them received four different treatments with the corresponding priority levels - L="Low", M="Medium", H="High", C="Critical". The table below shows the order of the treatments and the clinical result (change in heart rate) for each volunteer and treatment. The analysis of experiment includes diverse types of tests. Before running an experiment, a researcher must design a global plan, including the tests he wishes to use in the data analysis procedure after the test (Table 4&5).

In some circumstances, the preliminary analysis indicates that there may be some interesting results that cannot be analyzed through the preplanned trials.

Before proceeding with ANOVA analysis of LSD, we perform Box-and-Whisker d diagram, which is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile and maximum (Figure 3).

Table 4 Analysis of variance table for the Latin square design

Source of variation	Sum of squares	Degrees of freedom	Mean square	F ₀
Treatments	SS _{Treatments}	p-1	MS _{Treatments}	F ₀ =MS _{Treatment} /MS _E
Rows	SS _{Rows}	p-1	MS _{Rows}	
Columns	SS _{Columns}	p-1	MS _{Columns}	
Error	SS _E	(p-2)(p-1)	MS _E	
Total	SS _{Total}	p ² -1		

Table 5 Data for the clinical effect

	Position 1	Position 2	Position 3	Position 4
Volunteer 1	H=26.7	C=19.7	M=29	L=29.8
Volunteer 2	L=23.1	M=21.7	C=24.9	H=29
Volunteer 3	M=29.3	L=20.1	H=29	C=27.3
Volunteer 4	C=25.1	H=17.4	L=28.7	M=35.1

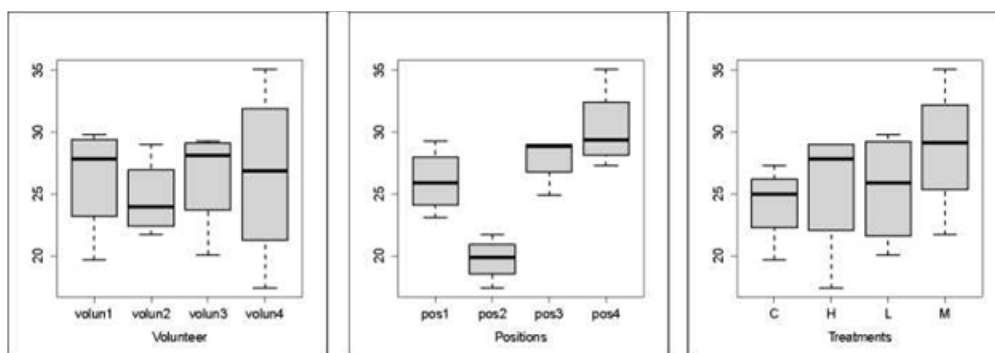


Figure 3 Box plots for Latin square design.

Note that the differences considering the volunteer is low, it is medium considering the treatments and high considering the positions. Now let confirm these graphic observations with the analysis of variance table. ANOVA is a set of statistical methods used mainly to compare the means of two or more samples. ANOVA can be treated as a special case of general linear regression where predictor variables are factors. Each value that can be taken by a factor is reflected to as a level. The build-in function in R *aov()* both examine a dependent variable and determine the variability of this variable in response to various factors. The results for the numerical example are listed in the table below.

Significance of *F* - test for null hypothesis: $H_0 : \tau_i = 0$ vs $H_a : \tau_i \neq 0$

- a) The difference between group considering the volunteer is **not significant (p-value>0.1)**.
- b) The difference between group considering the positions is **quite significant**.
- c) Among the treatments, there is no evidence of significant differences, at the significance levels of 1% and 5% (Table 6).

Table 6 Analysis of variance table for the clinical data

	Df	Sum of squares	Mean square	F value	Pr(>F)
Volunteer	3	9.427	3.142	0.8821	0.501548
Positions	3	245.912	81.971	23.0106	0,001084**
Treatments	3	45.277	15.092	4.2367	0,062818.
Residuals	6	21.374	3.562		

Note: *** significant at 0.1%, ** at 1%, * at 5%, . at 10%

Considerations and remarks

This paper explores the application of *BIBDs* and *LSDs* in statistical design of experiments. We revise the simplest combinatorial designs, as was previously stated, in order to summarize the basic idea of their usage. On the whole, the main reason to choose these designs is the opportunity to do a comparison between structures with one and two source of variation. Block designs provides error control measures for elimination in only one direction - block variations, whereas the improved design, Latin square can eliminate treatments effects using two source of variations, namely row and column.

As an extension of this work, we plan to consider particular cases of these combinatorial designs applying to other statistical models, exploring and improving the computational features in R.

Acknowledgements

Teresa A. Oliveira was partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal - FCT under the project UID/MAT/00006/2013.

Conflict of interests

Author declares that there is no conflict of interest.

References

1. Colbourn C, Dinitz J. *CRC Handbook of Combinatorial Designs*. Boca Raton: CRC Press; 2007.
2. Beth T, Jungnickel D, Lenz H. *Design Theory*. England: Cambridge University Press, Cambridge; 1999.
3. Hedayat AS, Sloane NJA, Stufken John. *Orthogonal arrays: theory and application*. Springer; 1999.
4. John, Peter William Meredith. *Incomplete block designs*. Marcel Dekker; 1980.
5. Montgomery, Douglas C. *Design and analysis of experiments*. 5th edn. John Wiley and Sons; 2001.
6. Oliveira TA. BIB designs with repeated blocks: review and perspectives. Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume I. The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan. 2010;82–96.
7. Brouwer AE. *The number of mutually orthogonal Latin squares—a table up to order 10000*. Report ZW 123. Amsterdam: Math Center; 1978.
8. Johnson DM, Dulmage AI, Mendelson NS. Orthomorphisms of group and orthogonal Latin squares. *Canadian J Math*. 1961;13:356–372.
9. Schellenberg PJ, Van Rees GM, Vanstone SA. Four pair wise orthogonal Latin squares of order 15. *Ars Combinatoria*. 1978;6:141–150.
10. Davim, Paulo J. *Statistical and Computational Techniques in Manufacturing*. Springer. 2012.
11. Todorov DT. Four mutually orthogonal Latin squares of order 14. *Journal of Combinatorial Designs*. 2012;20:1–5.
12. Abel RJR, Todorov DT. Four mutually orthogonal Latin squares of orders 20, 30, 38 and 44. *J Comb Theory(A)*. 1993;64:144–148.