JOURNAL OF ALGEBRAIC STATISTICS
Vol. 1, No. 1, 2010, 6-12
ISSN 1309-3452 – www.jalgstat.com

Abstract Algebra in Statistics

Wen-Fong Ke[1], Günter F. Pilz[2,*]

[1] *Department of Mathematics, National Cheng Kung University, and National Center for Theoretical Sciences (South), Tainan 701, Taiwan*
[2] *Department of Algebra, Johannes Kepler Universität Linz, Altenberger Strasse 69, 4040 Linz, Austria*

Abstract. In this note we show how some specific classes of algebraic structures ("planar near-rings") give rise to efficient Balanced Incomplete Block Designs, which in turn can excellently be used in statistical experiments.

2000 Mathematics Subject Classifications: 62K10, 05B05, 16Y30

Key Words and Phrases: BIB-Design, planar near-ring

---

1. The Algebraic-Combinatorial Settings

Definition 1. *A finite set N of size v, together with a collection $\mathcal{B}$ of b subsets of size k is called a* Balanced Block Design *if every element of N appears in a fixed number r of subsets of $\mathcal{B}$ and every pair of different elements of N is contained in the same number of λ subsets. We say that $\mathcal{B}$ is* incomplete *if it is not the set of all k-subsets of N, and, in this case, $(N, \mathcal{B})$ is then referred to as a* BIB-Design.*The elements of N are then called points, the sets in $\mathcal{B}$ are called blocks. The quintuple (v, b, r, k, λ) are the parameter of the design.*

Given a BIB-Design $(N, \mathcal{B})$, one can use it very well for statistical experiments. Suppose one wants to test combinations out of v "ingredients" (usually called "treatments") for a final product. Then one can start with any combination and vary only one ingredient at a time in a number of tests ("plots"). This is the traditional way, but it is a highly inefficient one. It is much better and saves a lot of costs to change several or all ingredients every time. It is a good ideas to do this in a "fair" manner: Each plot should get the same number of ingredients, each ingredient should get the same "chance" in that it is tested on the same number of plots, and each combination of different ingredients should be tested in the same number of plots (to get information on positive or negative synergy effects among the ingredients).

---

*Corresponding author.

Email address: wfke@mail.ncku.edu.tw (W.-F. Ke), guenter.pilz@jku.at (G. F. Pilz)

This can be achieved by choosing a (v, b, r, k, λ)-BIB Design; the ingredients are taken as points, the blocks as plots, and the elements in a block are just the ingredients which are applied in this plot. Then each of the b plots get precisely k ingredients, each ingredient is tested r times, and each pair of different ingredients come together in precisely λ plots. A concrete example follows below.

1. Tests of different ingredients to an optimal fertilizer on experimental agricultural (not algebraic) fields which are the plots.

2. The ingredients might be additives to paints to increase their resistance to sunlight, rain, etc. A plot is a test mixture for the paint.

3. For marketing purposes, one tests which combination of commercials work best to get higher sales of an article. The commercials (TV spots, sales actions, ...) are the ingredients in a particular test run (plot).

We want to show that one can easily get many (BIB-Designs) from a class of generalized rings, so-called *near-rings*.

Definition 2. *A set N, together with two operations $+$ and $\cdot$, is a near-ring provided that $(N, +)$ is a group (not necessarily abelian), $(N, \cdot)$ is a semigroup, and $(n + n') \cdot n'' = n \cdot n'' + n' \cdot n''$ holds for all $n, n', n'' \in N$.*

Of course, every ring is a near-ring; hence near-rings are generalized rings. Two ring axioms are missing: the commutativity of addition and (much more important) the other distributive law.

The standard example of a near-ring can be obtained by taking a group $(G, +)$, not necessarily abelian; then $G^G = \{f \mid f : G \to G\}$ is a near-ring with binary operations $+$ and $\circ$ given by $(f + g)(a) = f(a) + g(a)$ and $(f \circ g)(a) = f(g(a))$ for all $f, g \in G^G$ and $a \in G$. Furthermore, every near-ring can be embedded in some $(G^G, +, \circ)$ for some suitably chosen group G.

We need, however, a specific type of near-rings motivated by geometry.

Definition 3. *A near-ring N is called a planar near-ring if (1) there are at least two elements $a, b \in N$ such that $x \cdot a \neq 0$, $y \cdot b \neq 0$, and $z \cdot a \neq z \cdot b$ for some $x, y, z \in N$, and (2) all equations*

$$x \cdot a = x \cdot b + c, \quad (a, b, c \in N,\ z \cdot a \neq z \cdot b \textit{ for some } z \in N),$$

have exactly one solution $x \in N$.

This conditions of the planar near-rings mean that (1) there are at least two distinct non-zero "slopes", and (2) two "non-parallel" lines (described by some equations $y = x \cdot a + c_1$ and $y = x \cdot b + c_2$) have exactly one point of intersection.

There are various methods to construct planar near-rings (see [4] and [8]). Most of them use fixed-point free automorphism groups. Here we just present the easiest way which suits our need.

Definition 4 (Construction method for finite planar near-rings)**.** *Take a finite field $F = \mathrm{GF}(q)$, where q is a power of some prime p, and choose a generator g for its cyclic multiplicative group. Choose a proper factor t of $q-1$. Define a new multiplication $*_t$ in F as $g^m *_t g^n := g^{m+n-n_t}$, where $m, n \in \{1, 2, \ldots, q-1\}$, and $n_t \in \{0, 1, \ldots, t-1\}$ denotes the remainder of n on division by t; also set $0 *_t g^m = g^m *_t 0 = 0 \cdot 0 = 0$. Then $(F, +, *_t)$ is a planar near-ring.*

Example 1. *We give a (very small) example. Choose $F = \mathrm{GF}(7)$ with $g = 3$ a generator, and $t = 2$ as a divisor of $6 = 7-1$. From this, we get the multiplication table*

$*_2$	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	1	4	4	2
2	0	2	4	2	1	1	4
3	0	3	6	3	5	5	6
4	0	4	1	4	2	2	1
5	0	5	3	5	6	6	3
6	0	6	5	6	3	3	5

For example, in F we have $4 = 3^4$ (namely, $4 \equiv 3^4 \pmod 7$) and $3 = 3^1$, and the above table says that $4 *_2 3 = 3^4 *_2 3^1 = 3^{4+1-1} = 3^4 = 4$ in F.

2. The Experimental Settings

The last example can be readily used to design statistical experiments. Again, we show this via an example.

Example 2. *We want to test the combinations out of 7 ingredients for fertilizers. Testing all $2^7 = 128$ possible combinations of ingredients requires a huge amount of space and money. So we conduct an incomplete test. But this one should be* fair *to the ingredients (each ingredient should be applied the same number of times) and fair to the experimental fields (each test-field should get the same number of ingredients).*

We take the above near-ring of order 7 and form the sets $B_i = a *_2 N^* + b$, $a, b \in N$ with $a \neq 0$. Here N^* denotes the non-zero elements of N, and $a *_2 N^* = \{a *_2 x \mid x \in N, x \neq 0\}$ is abbreviated as aN^*:

$$\begin{array}{llll}
1N^* + 0 & = \{1,2,4\} =: B_1, & 3N^* + 0 & = \{3,5,6\} =: B_8, \\
1N^* + 1 & = \{2,3,5\} =: B_2, & 3N^* + 1 & = \{4,6,0\} =: B_9, \\
\vdots & & \vdots & \\
1N^* + 6 & = \{0,1,3\} =: B_7, & 3N^* + 6 & = \{2,4,5\} =: B_{14}.
\end{array}$$

(Notice that $1N^* = 2N^* = 4N^*$ and $3N^* = 5N^* = 6N^*$.) We see that these blocks form a BIB-design with $v = 7$ points (namely $0,1,2,3,4,5,6$) and $b = 14$ blocks $B_1, \ldots, B_{14}$ with each block contains precisely $k = 3$ elements; each point lies in exactly $r = 6$ blocks, and every pair of distinct points appears in $\lambda = 2$ blocks.

In order to solve our fertilizer problem, we divide the whole experimental area into 14 *experimental fields, which we number by* $1, 2, \ldots, 14$. *We then apply precisely the fertilizer* F_i *to a block* j *if* $i \in B_j$, $i = 0, 1, \ldots 6$. *Then we have to wait for the harvest time and we can measure the yields in the* 14 *blocks.*

Field / *Fertilizer*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
F_0				x		x	x		x	x		x		
F_1	x				x		x			x	x		x	
F_2	x	x				x					x	x		x
F_3		x	x				x	x				x	x	
F_4	x		x	x					x				x	x
F_5		x		x	x			x		x				x
F_6			x		x	x		x	x		x			
Yields y_i	*12.3*	*14.1*	*12.1*	*14.9*	*11.1*	*13.6*	*12.5*	*11.2*	*13.9*	*13.5*	*11.3*	*13.9*	*12.2*	*14.9*

The last row indicates the yields on the experimental fields after performing the experiment. We also include a 0^{th} *field on which we put nothing ("reference field"), because something will grow even if no fertilizer is put to the ground; on this field, we got a yield of* 10.5.

Then we get that every field contains exactly 3 *fertilizers and every fertilizer is applied to* 6 *fields. Moreover, every pair of different fertilizers is applied precisely twice in* direct competition. *That is, we have used a BIB-design with parameters* $(7, 14, 13, 3, 2)$.

The statistical analysis tries to give a "formula" for the yield y as $y = c+\beta_0 x_0+\beta_1 x_1+\cdots+\beta_{v-1}x_{v-1}$. We want to find best estimates for c and the βi and use the incidence matrix $\mathbb{A}$ of the design, which is a $(v+1)\times b$ matrix with 1 at the (i, j) position if the i^{th} element is in the j^{th} block, otherwise, 0 at that position. Take $\boldsymbol{y} = (y_1, y_2, \ldots, y_b)$, compute $(\mathbb{A}\mathbb{A}^t)^{-1}\mathbb{A}\boldsymbol{y}^t =: (c, \beta_0, \ldots, \beta_{v-1})^t$, and then Linear Algebra and/or Statistics tells us that $\beta_0, \ldots, \beta_{v-1}$ are the best estimates for the effects of the ingredients $F_0, \ldots, F_{v-1}$.

In our case, we get the best estimates for the effects of c and $F_i = \beta_i$, $i = 0, \ldots, 6$, as

$$(c, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6) = (10.5, 1.95, -0.4, 1.4, 0.4, 1.45, 1.3, -0.3).$$

Also, by the usual statistical analysis, we get confidence intervals for c and the effects of F_i. We see that β_1, β_3, and β_6 are not significant and thus leave off ingredients number 1, 3, and 6 and do the regression again. So we only apply F_0, F_2, F_4, F_5 to expect a yield of $10.3 + 2.0 + 1.5 + 1.5 + 1.4 = 16.7$, which is considerably better than the best yield of 14.9 which we got in the experiments. Other consideration may further improve the result (see Section 4 for more details).

3. A variety of designs from Planar Near-Rings

Given a finite planar near-ring $N = (\mathrm{GF}(q), +, *_t)$ from the construction in Definition 4, we can obtain several BIB-designs by choosing appropriately a collection of subsets of N

(called blocks). Recall that $a *_t N^*$ and $a *_t N$ are respectively abbreviated as aN^* and aN.

1. Take blocks the collection $\mathcal{B}^*$ of all $aN^* + b$ with $a, b \in N$ and $a \neq 0$. See [4, (5.5)].
2. Take blocks the collection $\mathcal{B}$ of all subsets of the form $aN + b$ with $a, b \in N$ and $a \neq 0$, provided that either all or none of the aN are additive subgroups (cases 1 or 2, respectively). See [4, (7.10) and (7.11)].
3. Take blocks the collection $\mathcal{B}^-$ of all $(aN \cup (-a)N) + b$ with $a, b \in N$ and $a \neq 0$, provided that for all nonzero a, $aN \cap (-a)N = \{0\}$, $aN \cup (-a)N$ is not an additive subgroup of N, and the map $t_a : N \to N; t_a(x) = x + ax$ is bijective. See [4, (7.99)].
4. Take blocks the collection $\mathcal{S}$ consisting of the intersection of $(b - a)N + a$ and $(a - b)N + b$ with $a \neq b$. See [9].

In general, similar results can be established for any finite planar near-rings not constructed from the method of Definition 4. For more details see [4] and [9].

4. Statistical Considerations

In the cases mentioned above, we get the following parameters for the corresponding experimental designs (recall that we use q for the cardinality of the near-ring N with underlying set $\mathrm{GF}(q)$, and we have $q - 1 = st$ with $s > 1$ and $t > 1$):

Theorem 1. *1. For $(N, \mathcal{B}^*)$, we can test q "ingredients"; for that, we need $b = qs$ tests with $k = t$ ingredients in each test. Every ingredient is tested $r = q - 1$ times, and each pair of different ingredients is tested $\lambda = t - 1$ times.*

2. In $(N, \mathcal{B})$, we need s^2 tests in case 1, in which we apply $k = t + 1$ ingredients in each test. Each ingredient is tested in exactly $r = s$ times, and each pair of different ingredients is tested $\lambda = 1$ time. For case 2, we need qs tests with again $t + 1$ ingredients in each test, but now $r = q + s$ and $\lambda = t + 1$.

3. In $(N, \mathcal{B}^-)$, we again get $b = qs/2$ tests with $k = 2t + 1$ ingredients in each test, $r = q + s/2$, and $\lambda = 2t + 1$. (Note that in this case, s is even.)

4. Finally, for $(N, \mathcal{S})$, $b = q(q - 1)/2$, some k and r (which can be determined by some equations), and $\lambda = k(k - 1)/2$.

We would like to remark that BIB-designs may have other properties besides the fairness in getting equal number of tests on any combination of different ingredients. There are BIB-designs having the same appearance but behaving differently from the structural point of view. That is, there are (in algebra sense) non-isomorphic BIB-designs having the same parameters. (See [6].) Also, there are BIB-designs such that any three distinct elements are contained in at most one block (cf. [3]), which may have some advantage over others that miss such characteristic. One may also want to take such structural considerations into account when choosing a BIB-design for experiments. The construction methods of Section 3 do produce such BIB-designs.

It is a good idea to care about possible (positive or negative) synergy effects. This is equivalent to asking if the linear model used here is really appropriate. The best way is to include, e.g., products of variables $x_i x_j$ into the linear model $y = c + \beta_0 x_0 + \cdots + \beta_{v-1} x_{v-1}$ and to test if these products $x_i x_j$ appear with significant coefficients in the regression result. The same might be done with terms like x_i^2, and so on.

Some practical advises on the analysis of experiments conducted on such BIB-designs might be useful:

1. One should try to plot all pairs $(x_{i,j}, y_j)$ for each experiment E_j $(1 \leq j \leq m)$:

$$(x_{1,j}, \ldots, x_{i,j} \ldots x_{n,j}; y_j)$$

   in order to see if a dependence between x_i and y is likely, and, if so, whether this relation seems to be linear or not.

2. Whether the BIB-Design model only allows to use an ingredient or not? If one wants to test an ingredient in, say, 3 levels (like using one, two, or three liters of it), one can use these levels as independent ingredients. So using 3 levels, one gets 2 more variables.

3. As a rule of thumb, the number b of experiments should at least be of the order of magnitude of the square of the number of variables (including the combinations like $x_i x_j$).

4. So one should avoid an "over-fitting" (too many variables). A good way to check how good a model might be is to look at the "P-value". This value should be very close to 0; something like 10^{-5} is usually not bad. The P-value decreases with the number of variables, so even better is, in most cases, that the R^2_{adj}-value which should be close to 1, because this value also takes into account how many variables were used. Wikipedia [10] gives a good account on this, for example.

In our example above, we might test the F_i for synergies. We find by inspection of the regression results that $F_2 * F_5$ has a positive synergy effect. We get estimates $(c, \beta_0, \beta_2, \beta_4, \beta_5, \beta_2 * \beta_5) = (10.6, 2.2, 0.8, 1.3, 0.7, 1.9)$ and so a total yield of 17.5, which is again much better. Also, the P- and the R^2_{adj}-values have developed nicely:

1. All variables: $P = 8 \times 10^{-4}$, $R^2_{adj} = 0.885$.

2. Variables x_0, x_2, x_4, x_5: $P = 1 \times 10^{-4}$, $R^2_{adj} = 0.843$.

3. Variables $x_0, x_2, x_4, x_5, x_2 \times x_5$: $P = 5 \times 10^{-7}$, $R^2_{adj} = 0.962$.

Even after all is said and done, one can not be sure that the model used was really accurate. But this is statistics, and one only can claim that the model was good with a certain probability. Observe that there does not exist a "best" model; this concept is not even well-defined. More on the analysis of experiments can, e.g., be found in [1], [2], [5] or [7].

Acknowledgement

The first author is supported by grant 97-2923-M-006-001-MY2 of the National Science Council, Taiwan, and the second author is supported by grant P19463 of the Austrian National Science Fund (FWF).

References

[1] R. A. Bailey. Design of Comparative Experiments. Cambridge University Press, Cambridge, 2008.

[2] G. Box, S. Hunter and W. G. Hunter. Statistics for Experimenters: Design, Innovation, and Discovery, 2nd Edition. Wiley, 2005.

[3] T. Boykett, W.-F. Ke and G. F. Pilz. *Multiple intersection property of field generated planar nearrings.* In preparation.

[4] J. R. Clay. Nearrings: Geneses and Applications. Oxford University Press, Oxford, 1992.

[5] D. R. Cox. Planning of Experiments. Wiley, 1992.

[6] W.-F. Ke. *On nonisomorphic BIBD with identical parameters.* Combinatorics '90 A. Barlotti et al. (Editors), pp. 337–346. Elsevier, 1992.

[7] J. Ledolter and A. J. Swersey. Testing 1-2-3. Stanford Business Books, Stanford, 2007.

[8] G. F. Pilz. Near-rings, Revised Edition. North-Holland, Amsterdam, 1983.

[9] H.-M. Sun. *Segments in a planar nearring.* Discrete Math. **240** (2001), 205–217.

[10] `http://en.wikipedia.org/wiki/Coefficient_of_determination`.