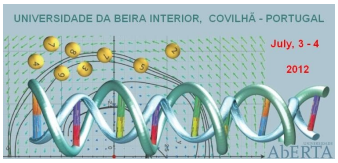
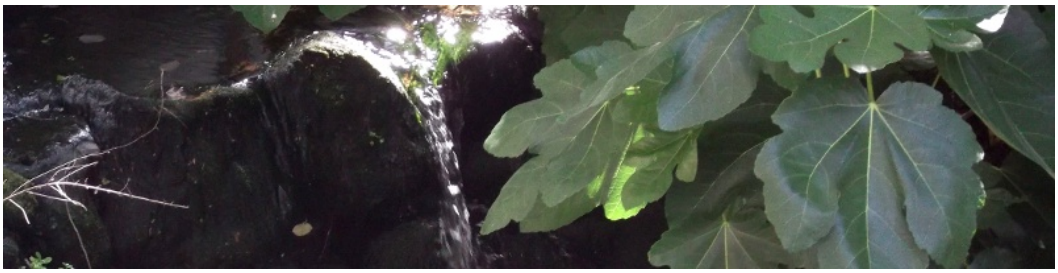
Balanced and variance-balanced designs

Peter J. Cameron
Queen Mary, University of London

6th Workshop on Statistics, Mathematics and Computation
Covilhã, Portugal, July 2012



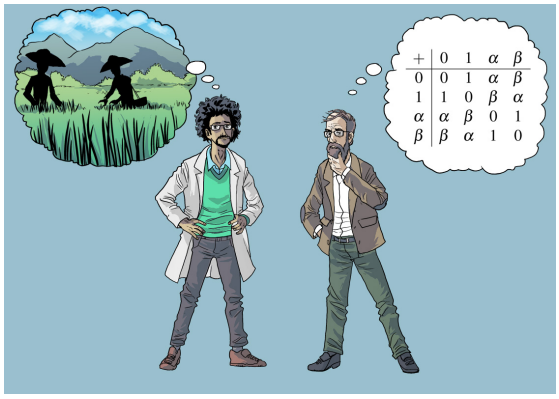


Mathematicians and statisticians

There is a very famous joke about Bose's work in Giridh. Professor Mahalanobis wanted Bose to visit the paddy fields and advise him on sampling problems for the estimation of yield of paddy. Bose did not very much like the idea, and he used to spend most of the time at home working on combinatorial problems using Galois fields. The workers of the ISI used to make a joke about this. Whenever Professor Mahalanobis asked about Bose, his secretary would say that Bose is working in fields, which kept the Professor happy.

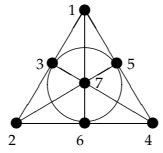
Bose memorial session, in *Sankhyā* **54** (1992) (special issue devoted to the memory of Raj Chandra Bose), i–viii.

Mathematicians and statisticians



Combinatorial and statistical design

In combinatorial design, a block is a set of points.
In statistical design, we have a set of plots; the treatments, and the blocks, each form a partition of the set of plots.



1	1	1	2	2	3	3
2	4	6	4	5	4	5
3	5	7	6	7	7	6

t-designs

A t-(v,k,λ) design consists of a set X of points and a collection $\mathcal{B}$ of blocks with the properties:

- there are v points;
- any block contains k points;
- any t points are contained in exactly λ blocks.

To avoid trivial cases, it is often assumed that $0 < t < k < v$ and $0 < \lambda$.
Thus, the Fano plane we met earlier is a 2-$(7,3,1)$ design.
For short, we often just refer to a t-design. Note that a 2-design is what statisticians call a balanced incomplete-block design or BIBD.
The main problem about t-designs is: for which values of the parameters do they exist?

Repeated blocks

The definition of a t-design allows the possibility that the same set of points occurs more than once as a block. In other words, we may have repeated blocks.
Thus, if $(X, \mathcal{B}_i)$ is a t-(v,k,λ_i) design for $i = 1, 2$, then $(X, \mathcal{B}_1 \cup \mathcal{B}_2)$ is a t-$(v,k,\lambda_1+\lambda_2)$ design.
In particular, there exists a 2-$(7,3,\lambda)$ design for all λ.
Some authors forbid repeated blocks, to make the existence question more challenging. (Clearly a 2-$(7,3,\lambda)$ design without repeated blocks is impossible for $\lambda > 5$.) There is no statistical reason for doing so.

Divisibility conditions

There are some well-known divisibility conditions for the existence of a t-design. Any such design is also an s design for $s \le t$, and the corresponding value of λ_s satisfies

$$\lambda_s \binom{k-s}{t-s} = \lambda \binom{v-s}{t-s},$$

where $\lambda = \lambda_t$; all the numbers λ_s must be integers, so $\binom{k-s}{t-s}$ divides $\binom{v-s}{t-s}\lambda$.
In particular, for a 2-design or BIBD, we have

- $k-1$ divides $(v-1)\lambda$;
- $k(k-1)$ divides $v(v-1)\lambda$;

where the quotients are the replication number $r = \lambda_1$ and the number of blocks $b = \lambda_0$.

The incidence matrix and the concurrence matrix

The incidence matrix of a design is the $v \times b$ matrix N whose rows are indexed by points and columns by blocks, with (x, B) entry 1 if $x \in B$ and 0 otherwise.
The matrix $NN^\top = \Lambda$ has rows and columns indexed by points, with (x, y) entry equal to the number of blocks containing x and y. This is the concurrence matrix of the design.
If our design is a 2-design, then the diagonal entries of Λ are equal to r (the replication number) and the off-diagonal entries to λ. So

$$NN^\top = (r - \lambda)I + \lambda J,$$

where J is the all-1 matrix.

Fisher's Inequality

Elementary operations on the matrix $(r - \lambda)I + \lambda J$ show that its determinant is equal to $(r + \lambda(v - 1))(r - \lambda)^{v-1} = rk(r - \lambda)^{v-1}$.
By our assumptions, the determinant is non-zero.
So the incidence matrix N has rank v, and so necessarily $b \geq v$.
This is Fisher's Inequality, another necessary condition for the existence of a 2-design (or indeed, of a t-design with $t \geq 2$.
Further manipulation shows that equality holds if and only if any two blocks intersect in a constant number of points (and this constant is necessarily λ), so that the dual design (obtained by interchanging the roles of points and blocks) is also a 2-design.
Such designs have been given many different names ("symmetric", "projective"); my preferred term is square 2-designs (since the incidence matrix is square).

Wilson's theorem

For 2-designs, there is a very satisfactory asymptotic existence theorem, due to Richard Wilson. No analogous result is known for t-designs with $t > 2$.

Theorem
Let k and λ be positive integers. Then for all sufficiently large v satisying the divisibility conditions $(k - 1) \mid (v - 1)\lambda$ and $k(k - 1) \mid v(v - 1)\lambda$, a 2-$(v, k, \lambda)$ design exists.

"Sufficiently large" is exponential, but certainly not large when compared to some of the functions arising in modern combinatorics!

Block size 3

The divisibility conditions for 2-$(v, 3, 1)$ designs read as follows:

- if $\lambda \equiv 1$ or 5 (mod 6), then $v \equiv 1$ or 3 (mod 6);
- if $\lambda \equiv 2$ or 4 (mod 6), then $v \equiv 0$ or 1 (mod 3);
- if $\lambda \equiv 3$ (mod 6), then v is odd;
- if $\lambda \equiv 0$ (mod 6), then no restriction on v.

These conditions are also sufficient. Indeed, many designs exist. For example, for $\lambda = 1$, these designs are Steiner triple systems, whose existence was established by Kirkman in the nineteenth century; it is known that there are at least $(cv)^{v^2/6}$ non-isomorphic designs on v points if v satisfies the necessary conditions, where $c > 0$.

A Markov chain

Since there are too many designs to list them all, we would like to explore them by some kind of random walk. This method is based on one for Latin squares due to Jacobson and Matthews.
Let us reformulate the definition of a 2-$(v, 3, \lambda)$ design. We assign a non-negative integer multiplicity to every 3-element set of points, in such a way that the sum of the multiplicities of all the sets containing any given pair of points is equal to λ.
Now an improper design is defined similarly, but with one 3-set having a multiplicity of -1 and all other multiplicies non-negative, and satisfying the same constraint for all point pairs.

Now the state space of the Markov chain consists of all proper and improper 2-$(v, 3, \lambda)$ designs with fixed λ. One move in the chain is defined as follows:

- Choose a 3-set abc whose multiplicity we want to increase by 1. If the design is proper, this may be any 3-set whose multiplicity is smaller than λ; if it is improper, it must be the 3-set with multiplicity -1.
- Choose any 3-sets $a'bc$, $ab'c$ and abc' with positive multiplicity (note that these exist).
- Increase by 1 the multiplicities of abc, $ab'c'$, $a'bc'$ and $a'b'c$, and decrease by 1 the multiplicities of $a'bc$, $ab'c$, abc' and $a'b'c'$. [This move is known as a trade].

If the multiplicity of $a'b'c'$ was positive, we obtain a proper design; otherwise we obtain an improper design with $a'b'c'$ as the negative block.

Conjecture

If v and λ are such that 2-$(v,3,\lambda)$ designs exist, then the Markov chain just described is connected.

If this is true, and if the choices in the theorem are made with appropriate probabilities (i.e., in the first step the 3-set is chosen with probability proportional to λ minus its multiplicity, and in the second step the choices are made uniformly), then the limiting distribution of the Markov chain has all proper designs equally likely (and all improper designs equally likely).
So if we wander for long enough, and then choose the first proper design we meet, the result will approach the uniform distribution on proper designs.

An example

We consider the case $v = 9$, $\lambda = 1$ (Steiner triple systems on 9 points). There is a unique such system up to isomorphism; its blocks are

123	147	159	357
456	258	267	249
789	369	348	168

At the first step we choose to add the block 124. We must remove the blocks 123, 174, 924, add in the blocks 173, 923, 974, and remove the block 973; since it is not a block, it ends up with multiplicity -1.
At the next step, we must put back 973, and can remove one of 974 and 978, one of 923 or 963, and one of 173 or 573. Suppose we choose 978, 923 and 573. Then we add 928, 578 and 523, and remove 528. Since it is a block, we get a proper system, with blocks

124	137	159	168
369	456	267	479
578	289	348	235

Binary and non-binary

Recall that a design is binary if no two treatments occur together in a block. In a non-binary design, a block has to be regarded a a multiset (rather than a set) of points or treatments: that is, each point has a multiplicity in the block.
For example,

1	1	1	1	2	2	2
1	3	3	4	3	3	4
2	4	5	5	4	5	5

The first block is the multiset $[1,1,2]$ containing two occurrences of 1 and one of 2. The other blocks are sets of three treatments.
This design is in fact E-optimal, that is, it minimizes the maximum variance of a normalized contrast over all block designs with $v = 5$, $b = 7$, $k = 3$.

Incidence matrix and concurrence matrix again

The incidence matrix has to be re-defined for non-binary designs: as before, rows are indexed by points, and columns by blocks; the (x,B) entry is the multiplicity of point x in block B.
The concurrence matrix is (as before) given by $\Lambda = NN^\top$. Its (x,y) entry is "the number of occurrences of x and y in the same block", that is, the number of ordered pairs of plots in the same block such that the first gets treatment x and the second gets treatment y.
In our example, $\Lambda_{12} = 2$, since each of the two occurrences of 1 in the first block can be paired with the unique occurrence of 2.

Variance-balanced designs

1	1	1	1	2	2	2
1	3	3	4	3	3	4
2	4	5	5	4	5	5

Note that any two treatments have two occurrences in the same block, so all off-diagonal elements of Λ are equal to 2. On the other hand, $\Lambda_{11} = 7$, whereas $\Lambda_{xx} = 4$ for $x \neq 1$.
A block design is called variance-balanced if all off-diagonal elements of the concurrence matrix Λ are equal. Thus a binary variance-balanced design is the same thing as a BIBD or 2-design.
We use the notation VB(v,k,λ). Thus the above design is a $VB(5,3,2)$.

An example

Consider variance-balanced designs with $v = b = 7$, $k = 6$.
Here are two examples:

- the design whose blocks are all the 6-subsets of the set of points;
- the design obtained from the Fano plane by doubling each occurrence of a point in a block (so that the first block is the multiset $[1,1,2,2,3,3]$.

The first is a BIBD with $\lambda = 5$. The second has $\lambda = 4$. Not surprisingly, the first is "better" in most reasonable senses.

Maximum trace

Morgan and Srivastav showed that a variance-balanced design with "not too much non-binarity" is E-optimal. I will briefly outline their condition, but have no time to explain it further. They define two new parameters of a VB design, as follows:

$$r = \left\lfloor \frac{bk}{v} \right\rfloor, \qquad p = bk - vr,$$

so that $bk = vr + p$ and $0 \leq p \leq v - 1$. Thus, in a BIBD we have $p = 0$. Note that the use of r does not here imply that the design has constant replication!
They further say that a VB design has maximal trace if its parameters satisfy the equation $r(k-1) = (v-1)\lambda$.
In our examples above, $r = \lfloor 7 \cdot 6/7 \rfloor = 6$ and $p = 0$. Since $r(k-1)/(v-1) = 6 \cdot 5/6 = 5$, we see that the first design has maximal trace, but the second does not.

The existence question

The main question about variance-balanced designs is:

Problem
For which values of the parameters v, k, λ does a $\mathrm{VB}(v,k,\lambda)$ of maximal trace exist? More generally, what is the minimum number of blocks in such a design?

Morgan and Srivastav answered the question for block size 3, as I now describe briefly. Note that for a $\mathrm{VB}(7,3,2)$, we can take any collection of 3-sets covering each pair of points 0 or 2 times, and then "fill up" with blocks of the form $[x,x,y]$ for uncovered pairs (x,y). As an exercise, construct such designs with 7, 8, 9 or 10 blocks.

Block size 3

In a $\mathrm{VB}(v,3,\lambda)$ which is not a BIBD, there must be a non-binary block $[x,x,y]$, so $\lambda \geq 2$. Also, if the design has maximal trace, then $2r = (v-1)\lambda$, so either v is odd or λ is even.

Theorem
A $\mathrm{VB}(v,3,\lambda)$ design of maximal trace exists whenever $\lambda(v-1)$ is even and $\lambda > 1$.

Indeed, there are many such designs (as with Steiner triple systems).

Problem
Is there a Markov chain method of exploring these designs?