Fisher and Bose, Hamming and Golay

Peter J Cameron
School of Mathematical Sciences
Queen Mary and Westfield College
Londn E1 4NS
p.j.cameron@qmw.ac.uk

Hamming codes

R. A. Fisher, The theory of confounding in factorial experiments in relation to the theory of groups, *Ann. Eugenics* **11** (1942), 341–353.

R. A. Fisher, A system of confounding for factors with more than two alternatives, giving completely orthogonal cubes and higher powers, *Ann. Eugenics* **12** (1945), 2283–290.

M. J. E. Golay, Notes on digital coding, *Proc. IEEE* **37** (1949), 657.

R. W. Hamming, Error detecting and error correcting codes, *Bell Systems Tech. J.* **29** (1950), 147–160.

Coding theory

We wish to send words of length n over an alphabet A with $|A| = q$ over a noisy channel where errors can occur.

We *assume* that, with high probability, not too many errors occur during transmission of a word.

The strategy is to send words from a *code* C, a subset of A^n. We require:

(a) *large minimum distance* d: if $d \geq 2e+1$, we can correct up to e errors;

(b) *many codewords* (subject to (a)): the transmission rate is $\log_q |C|/n$;

(c) *computationally efficient encoding and decoding* (subject to (a) and (b)).

Factorial design

We are investigating n factors which can affect the yield of some process. The ith factor can take any one of a set A_i of levels, with $|A_i| = q_i$.

We assume that only the interactions of small numbers of factors affect the yield significantly.

We impose the structure of an abelian group on A_i, and test treatment combinations lying in a subgroup B of $A_1 \times \cdots \times A_n$.

Factorial design

Let C be the annihilator of B in $A_1^* \times \cdots \times A_n^*$. (Here A_i^* is the group of characters of A_i; so C is the set of all characters of $A_i \times \ldots \times A_n$ which are trivial on B.)

Elements of C represent combinations of treatments which are confounded in the experiment. (For example, if an element of C has support in $A_i^* \cup A_j^* \cup A_k^*$, then the interaction of factors j and k cannot be distinguished from the main effect of factor i.)

Factorial design

We want

(a) *Large weight in* C so that potentially significant combinations of factors are not confounded;

(b) *Few trials* (subject to (a)): trials are expensive! This means small B, and so large C: note that

$$|C| = \frac{q_1 \cdots q_n}{|B|}.$$

(c) *simple description* which can be explained to experimenters and for which results can be analysed (subject to (a) and (b)).

Comparison

Design theorists and coding theorists are both looking for subsets C of $A_1 \times \cdots \times A_n$ with large minimum distance and large cardinality.

Coding theorists have n large, all A_i of the same size (almost always 2), and don't insist on group structure (though it does help to use a linear code).

Statisticians have n fairly small, varying alphabet size, and do require group structure.

Hamming codes satisfy both specifications!

Hamming codes

Let $V = \mathrm{GF}(q)^k$. Partition the non-zero vectors in V into equivlence classes, where two vectors are equivalent if one is a non-zero scalar multiple of the other. There are $(q^k - 1)/(q - 1)$ equivalence classes.

Choose one vector from each equivalence class, and let H be the $k \times (q^k - 1)/(q - 1)$ matrix having these vectors as columns. (For simplicity, take all vectors whose first non-zero entry is 1.) Then any two columns of H are linearly independent.

The code C with parity check matrix H thus has minimum weight 3 and so is 1-error-correcting. This is the *Hamming code* $H(k,q)$.

Fisher's Theorem on Minimal Confounding

Fisher (1942) proved that:

> A 2^n factorial scheme can be arranged in 2^{n-p} blocks of 2^p plots each, without confounding either main effects or 2-factor interactions, provided that $n < 2^p$.

Subsequently (1945), he generalized this theorem and proved that:

> A π^n factorial scheme can be arranged in π^{n-p} blocks of π^p plots each, without confounding either main effects or 2-factor interactions, provided that
>
> $$n \leq (\pi^p - 1)/(\pi - 1).$$

D. J. Finney, *An Introduction to the Theory of Experimental Design*, University of Chicago Press, Chicago, 1960.

(Here π is a prime power.)

Coding theory with mixed alphabets

C is a code of length n and minimum distance d over alphabets of size $q_1, \ldots, q_n$. Let $e = \lfloor (d-1)/2 \rfloor$, and assume that $q_1 \le \cdots \le q_n$.

Sphere-packing bound:

$$|C| \le \frac{\prod_{i=1}^{n} q_i}{\sum_{k=0}^{e} \sum_{i_1 < \cdots < i_k} \prod_{j=1}^{k} (q_{i_j} - 1)}.$$

Singleton bound:

$$|C| \le \prod_{i=1}^{n-d+1} q_i.$$

Plotkin bound: Let

$$\alpha = \sum_{i=1}^{n} (1 - 1/q_i).$$

If $d > \alpha$ then $|C| \le d/(d-\alpha)$.

An example

Let $n = 5$ and let the alphabet sizes be $2, 2, 2, 2, 4$. Take $d = 3$.

The sphere-packing bound gives

$$|C| \leq \frac{2 \cdot 2 \cdot 2 \cdot 2 \cdot 4}{1+1+1+1+1+3} = 8.$$

The Singleton bound gives

$$|C| \leq 2 \cdot 2 \cdot 2 = 8.$$

The Plotkin bound:

$$\alpha = \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{3}{4} = \tfrac{11}{4} < 3,$$

so $|C| \leq 3/(3 - \frac{11}{4}) = 12$.

An example

Take $A_1 = \ldots = A_4 = \{0,1\}$ (the cyclic group of order 2) and $A_5 = \{0,a,b,c\}$ with $a+b+c=0$ (the Klein group of order 4).

Then C is

$$\begin{array}{c} 00000 \\ 11110 \\ 0011a \\ 1100a \\ 0101b \\ 1010b \\ 0110c \\ 1001c \end{array}$$

Codes and projective spaces

R. C. Bose, Mathematical theory of the symmetrical factorial design, *Sankhyā* **8** (1947), 107–166.

R. C. Bose and J. N. Srivastava, On a bound useful in the theory of factorial design and error-correcting codes, *Ann. Math. Statist.* **35** (1964), 408–414.

C. Greene, Weight enumeration and the geometry of linear codes, *Studies in Applied Math.* **55** (1976), 119–128.

Codes and projective spaces

Let A be a $k \times n$ matrix over $\mathrm{GF}(q)$. Assume that no two columns are linearly dependent, and that A has rank k.

(a) A is the parity check matrix of a $[n, n-k]$ code

$$C = \{v \in \mathrm{GF}(q)^n : Av^\top = 0\}.$$

Elementary row operations don't affect C; column permutations and scalar multiplications replace it by an equivalent code (metric properties are unaffected). The code C has minimum weight at least 3, so is 1-error-correcting. The corresponding factorial design has q^k treatments.

Codes and projective spaces

(b) The columns of A are a set S of n points in projective space $\mathrm{PG}(k-1,q)$. Elementary row operations induce collineations of the projective space, while column permutations don't change S. The set S spans $\mathrm{PG}(k-1,q)$.

So 1-error-correcting codes (up to equivalence) correspond naturally to spanning subsets of projective space (up to collineations).

The correspondence between codes and projective spaces allows many properties to be transferred back and forth:

Codes and projective spaces

1. The Hamming codes correspond to the entire projective space. The code/projective space connection can be regarded as a generalisation of the construction of Hamming codes.

2. Supports of words of the dual code correspond to complements of hyperplane sections of S.

3. (Bose 1947) MDS codes (those which meet the Singleton bound) correspond to arcs in projective space. (This, and a bound on the size of arcs in projective planes, are in Bose's paper on factorial designs.)

4. (Greene 1976) The weight enumerator of the code is a specialisation of the Tutte polynomial of the matroid represented by the matrix. Hence the MacWilliams identities follow from matroid duality.

An application

We are given a set of n objects, containing one ‘active pair’.

We can test any subset: the test is positive precisely when the subset contains both members of the active pair.

How many tests are required to identify the active pair?

(This problem arises in PCR tests in genetics: I learned about it from G. Gutin.)

An application

Suppose that $n = 2^d - 1$. Let H be the $2d \times n$ parity check matrix of a 2-error-correcting BCH code.

For each row of H, test the sets of positions where 0s occur and where 1s occur in that row. From these tests we can determine the syndrome and hence the active pair.

The number of tests is $4d$, which is just twice the information-theoretic lower bound. (And, if we get a positive result from a subset, we don't have to test the complementary subset.)