5-12-1996

CWATSETS: Weights, Cardinalities, and Generalizations

Richard Mohr
Rose-Hulman Institute of Technology

CWATSETS: Weights, Cardinalities, and Generalizations

Rick Mohr

MS TR 96-03

May 12, 1996

Department of Mathematics
Rose-Hulman Institute of Technology
Terre Haute, IN 47803

FAX (812) 877-3198 **Phone: (812) 877-839**

CWATSETS: Weights, Cardinalities, and Generalizations

Rick Mohr*

May 12, 1996

Abstract

This report provides an upper bound on the average weight of an element in a cwatset and discusses the ratio of the cardinality of a cwatset to the cardinality of the group containing the cwatset. The concept of a generalized cwatset is also introduced.

1 Introduction

Cwatsets are structures that are contained in binary space which are not quite a group but are more structured than just a set. The recent development of cwatsets has found uses in determining confidence intervals in statistics.

Definition 1 *A subset,* C, *of* $\mathbf{Z}_2^d$ *(binary d-space) with order* n *is a cwatset of degree* d *and order* n *if, for each element* $\mathbf{b}$ *of* C, *there exists* σ *in* S_d *(the symmetric group on d symbols) such that* C $+\ \mathbf{b} = C^\sigma$.

All cwatsets in $\mathbf{Z}_2^d$ can be found using projections of the subgroups of the wreath product of $\mathbf{Z}_2$ by S_d (Sherman and Wattenberg [2]). Kerr [3] used this approach to construct all cwatsets of degrees 3, 4, and 5.

*Author supported by NSF grant DMS-9322338

There are still many questions involved with the structure of cwatsets. Problems discussed in this paper include computing the average weight of a cwatset element, determining cwatsets of maximum order, and generalizing the definition of a cwatset to an arbitrary group.

2 Notation

There are a few notational ideas concerning cwatsets that will be used throughout this paper.

1) $w(x)$ will denote the weight enumerator of a cwatset, and $\bar{w}$ will be used to represent the average weight of a word for a cwatset.

2) A cwatset can have 'columns' if it is written so that the elements of the cwatset appear one above the other, similar to a matrix. For example, the six element cwatset below has four columns:

0000

0011

1001

0111

0100

1101

3 Average Weight

Sherman and Wattenberg [2] proved several facts about cwatsets which involved the weights of certain elements in the cwatset:

Fact 1 *A cwatset contains* **0** *(i.e. - a word of* weight *zero).*

Fact 2 *If a cwatset contains* **1** *or an element of odd* weight*, then the order of the cwatset is even.*

Since the average weight of an element is simply the total number of 1's in the cwatset divided by the order of the cwatset,

$$
\begin{aligned}
n \cdot \bar{w} &= a_1 + 2a_2 + \cdots + (n-1)a_{n-1} \\
&= [a_1 + (n-1)a_{n-1}] + [a_2 + (n-2)a_{n-2}] + \cdots + \frac{n}{2}a_{n/2} \\
&\leq [\frac{n}{2}(a_1 + a_{n-1})] + [\frac{n}{2}(a_2 + a_{n-2})] + \cdots + \frac{n}{2}a_{n/2} \\
&\leq \frac{n}{2}(a_0 + a_1 + a_2 + \cdots + a_{n-1}) \\
&= \frac{n \cdot d}{2}
\end{aligned}
$$

Thus

$$\bar{w} \leq \tfrac{d}{2} \quad \square$$

It's important to note a few things about the proof. First, a_n is not included because 0 is always in C, which makes $a_n = 0$. Also, the term $a_{n/2}$ was included, establishing the result if n is even. If n is odd, this term is omitted, and the proof still holds. Finally, the average weight is exactly half the degree only if all columns have the same weight (i.e. - the only non-zero term is $a_{n/2}$.)

It is known that for a subgroup of $\mathbf{Z}_2^d$, the average weight of an element is exactly d/2 (provided there are no columns of zeros). If the subgroup is written so that we can refer to columns, then the number of 1's in each column is exactly n/2. Thus, the result about subgoups becomes a corollary to this theorem.

Some observations about this theorem:

1) It can be used to eliminate subsets of $\mathbf{Z}_2^d$ as candidates for cwatsets. Simply compute the average weight of the subset, and if it is greater than d/2, the subset cannot possibly be a cwatset (the catch is, if $\bar{w} \leq$ d/2, this doesn't prove the subset is a cwatset.)

2) It provides a possible justification of the use of half samples ([4] and [5]) in statistics. It was proved in [1] that cwatsets can be used to create equal probability confidence intervals. Now, the

Fact 3 *The elements of even* **weight** *in a cwatset form a (sub)cwatset.*

Fact 4 *Let C be a cwatset in $\mathbf{Z}_2^d$ of order* n *such that* n $> 2^{d-1}$. *Then, C contains an element of* **weight** *k, for k = 0,1,...,d.*

A natural question to ask: On the average, what is the weight of an element in a cwatset?

To provide an upper bound, we will use the following lemma.

Lemma 1 *If a_k is the number of columns of weight k in a cwatset of order n, $k \cdot a_k = (n - k) \cdot a_{n-k}$.*

This lemma is a generalization of a lemma given by Kerr [3] and can be proven with the same proof she uses.

We can prove the following theorem [1] using Lemma 1.

Theorem 1 *The average weight of an element of a cwatset of degree d is less than or equal to $d/2$.*

Proof. Without loss of generality, we will assume $k \leq \frac{n}{2}$. From Lemma 1, we see that $a_{n-k} \leq a_k$ with equality only if $k = \frac{n}{2}$ or $a_k = a_{n-k} = 0$. This gives us

$$a_{n-k} \leq a_k$$

$$na_{n-k} - 2ka_{n-k} \leq na_k - 2ka_k$$

$$2na_{n-k} - 2ka_{n-k} \leq na_k - 2ka_k + na_{n-k}$$

$$2ka_k + 2na_{n-k} - 2ka_{n-k} \leq na_k + na_{n-k}$$

$$ka_k + (n-k)a_{n-k} \leq \frac{1}{2}n(a_k + a_{n-k})$$

[1] A proof of this theorem using dot products of cwatset elements is given in appendix A.

average subsample size of the cwatset elements is about half the size of the total sample. So, on the average, cwatsets act like half samples. It is known that cwatsets give good approximations to the underlying mean, so its seems likely that half samples should do well also.

The average weight of a cwatset can also be computed using the weight enumerator. If $w(x)$ is the weight enumerator, then $\bar{w} = \frac{w'(1)}{w(1)}$. With this, we can prove the following fact.

Fact 5 *Let C_1 and C_2 be cwatsets with average weights $\bar{w_1}$ and $\bar{w_2}$ respectively. Then if $C_3 = C_1 \oplus C_2$, the average weight of C_3 is $\bar{w_3} = \bar{w_1} + \bar{w_2}$.*

Proof. If $w_i(x)$ is the weight enumerator for C_i, then $w_3(x) = w_1(x) \cdot w_2(x)$.[2] So,

$$\begin{aligned}
\bar{w_3} &= \frac{w_3'(1)}{w_3(1)} \\
&= \frac{w_1'(1)w_2(1) + w_1(1)w_2'(1)}{w_1(1)w_2(1)} \\
&= \frac{w_1'(1)}{w_1(1)} + \frac{w_2'(1)}{w_2(1)} \\
&= \bar{w_1} + \bar{w_2}. \quad \square
\end{aligned}$$

4 Maximal Order Cwatsets?

Another problem to ponder which is linked to the use of cwatsets in statistics: What is the proper cwatset of maximal order contained in $\mathbf{Z}_2^d$ for a given value of d?

To make discussing this problem a little easier, we'll introduce some notation. Let $\hat{C_d}$ be the maximal order proper cwatset in $\mathbf{Z}_2^d$. If there is no unique maximal order proper cwatset, then choose $\hat{C_d}$ to be any one of these cwatsets. Next, we introduce the following definition.

[2]This result can also be used as a technique for decomposing cwatsets (see appendix B).

Definition 2 *For any cwatset, C, of degeree* d *and order* n, *let*

$$\rho(C) = \frac{n}{2^d}$$

.

Now instead of looking at the order of the largest cwatset, we can look at $\rho(\hat{C}_d)$. From the knowledge of existing cwatsets [3], the values of $\rho(\hat{C}_d)$ can be computed for some of the lower dimensions.

d	1	2	3	4	5	6	7
ρ	1/2	1/2	3/4	3/4	3/4	3/4 (?)	7/8 (?)

The values of ρ for dimensions 1 through 5 are exact. All evidence suggests the values for dimensions 6 and 7 are correct; however, the possibility of higher ratios has not been disproven. Even though the ratios appear to be following a nice increasing pattern, there is still no proof as to what the value of ρ is for a given dimension or even that this pattern continues.

The question now becomes finding out what the limit of $\rho(\hat{C}_d)$ is. Since the order of $\hat{C}_{d+1}$ must be greater than or equal to the order of $\{0,1\} \oplus \hat{C}_d$, we know that the ratios must be non-decreasing. But does the limit go to 1 (meaning the cardinality of a cwatset can be arbitrarily close to the cardinality of the entire group), or does it approach some other limit (and if so, what is it)?

5 Generalized Cwatsets

Definition 3 *Let $H_\phi \times G$ be the semidirect product of groups H and G determined by the homomorphism $\phi : H \mapsto Aut(G)$. A subset, C, of G is defined to be an (H, ϕ) subset of G if for every* **b** $\in C$ *there exists $h \in H$ such that*

$$C \cdot \mathbf{b} = C^{\phi(h)}$$

where $(\cdot)$ is the corresponding binary operation of the group.

Notice that if $G = \mathbf{Z}_2^d$, $H = S_d$, and ϕ is the appropriate homomorphism, then we get our defintion of a cwatset. So (H, ϕ) subsets are really a generalization of the 'classical' cwatset definition. In fact, when G, H,and ϕ are clearly determined, we will refer to C as a cwatset of G (because our set is still closed with a twist).

Fact 6 *Let C be an (H, ϕ) subset of the group G. The identity element, e, of G is contained in C if and only if for every x in C, x^{-1} is also in C.*

Proof. Assume $e \in C$. Then $e \in C^{\phi(h)}$ since $e^{\phi(h)} = e$ for any automorphism. For any $x \in C$, $C \cdot x$ contains e, so there must exist $y \in C$ such that $y \cdot x = e$. This implies $y = x^{-1} \in C$.

Now assume that for every $x \in C$, $x^{-1} \in C$. Thus $e = x^{-1} \cdot x \in C \cdot x = C^{\phi(h)}$. So there exists $y \in C$ such that $y^{\phi(h)} = e$. But this is only possible if $y = e$. □

This result should come as no surprise. It is simply a generalization of Fact 1, because in a classical cwatest each element is also its own inverse.

In the classical context, all cwatsets are projections of subgroups of $S_{d_\phi} \times \mathbf{Z}_2^d$. This means that the order of a cwatset always divides $2^d \cdot d! = |S_{d_\phi} \times \mathbf{Z}_2^d|$ (Sherman and Wattenberg [2]). However, in (H, ϕ) subsets, we run into trouble. Projections of subgroups of $H_\phi \times G$ are still cwatsets, but now there are cwatsets which do not contain the identity element, meaning they are not projections of subgroups. So does the order of an (H, ϕ) subset still divide $|H_\phi \times G|$? If the cwatset contains the identity, then the answer is yes (just use the same argument that was used for classical cwatests [2]). In other cases, it also appears to be true, but there is still no proof. I conjecture that the order of an (H, ϕ) subset must divide $|H_\phi \times G|$.

6 (H, ϕ) subsets of $\mathbf{Z}_2^d$

If $G = \mathbf{Z}_2^d$ and $H = \text{Aut}(\mathbf{Z}_2^d)$, instead of just S_d, then we can solve the problem of finding the maximal order cwatsets in a given dimension. In fact, it's easy to show that

$$\rho(\hat{C}_d) = \frac{2^d - 1}{2^d}$$

The basic idea is that, since we have more automorphisms to use now, we can form a cwatset just by throwing out any non-zero word from the group. This solves the earlier problem in the sense that we now know what ratio occurs for each dimension and that the ratios go to 1 in the limit of big d. So, we can get a cwatset with an order arbitrarily close to the order of the entire group. Even better, we can get the ratio of the order of a cwatest to the order of the group arbitrarily close to any fraction, as we will soon see.

Lemma 2 $\rho(C_1 \oplus C_2) = \rho(C_1)\ \rho(C_2)$.

Proof. Let C_1 and C_2 have dimensions a and b respectively. Then

$$\rho(C_1 \oplus C_2) = \frac{|C_1 \oplus C_2|}{2^{a+b}} = \frac{|C_1| \cdot |C_2|}{2^a 2^b} = \rho(C_1)\ \rho(C_2) \quad \square$$

Theorem 2 *There exists a sequence of cwatsets* $\{C_n\}$ *such that* $\lim_{n \to \infty} \rho(C_n) = r$ *for* $0 \leq r \leq 1$.

Proof. The idea behind constructing the sequence of cwatsets is to approximate r by a sequence of products of numbers of the form $\frac{2^d - 1}{2^d}$ (1/2, 3/4, 7/8, etc.). For example, suppose $r = 0.317$. Then we can form a sequence with a limit of r.

$$r_1 = \frac{1}{2} = 0.5$$

$$r_2 = \frac{1}{2} \cdot \frac{3}{4} = 0.375$$

$$r_3 = \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{7}{8} = 0.328125$$

$$r_4 = \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{7}{8} \cdot \frac{31}{32} = 0.31787109375$$

...etc.

Now we use Lemma 2 and the fact that

$$\rho(\hat{C}_d) = \frac{2^d - 1}{2^d}$$

to construct the sequence of cwatsets needed. For our example,

$$C_1 = \hat{C}_1$$

$$C_2 = \hat{C}_1 \oplus \hat{C}_2$$

$$C_3 = \hat{C}_1 \oplus \hat{C}_2 \oplus \hat{C}_3$$

$$C_4 = \hat{C}_1 \oplus \hat{C}_2 \oplus \hat{C}_3 \oplus \hat{C}_5$$

...etc.

The reason for choosing these cwatsets as the sequence is seen below.

$$\rho(C_1) = \rho(\hat{C}_1) = \frac{1}{2} = 0.5 = r_1$$

$$\rho(C_2) = \rho(\hat{C}_1 \oplus \hat{C}_2) = \rho(\hat{C}_1)\rho(\hat{C}_2) = \frac{1}{2} \cdot \frac{3}{4} = 0.375 = r_2$$

$$\rho(C_3) = \rho(\hat{C}_1 \oplus \hat{C}_2 \oplus \hat{C}_3) = \rho(\hat{C}_1)\rho(\hat{C}_2)\rho(\hat{C}_3) = \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{7}{8} = 0.328125 = r_3$$

Thus,

$$\lim_{n\to\infty} \rho(C_n) = \lim_{n\to\infty} r_n = r$$

This same argument holds for any $0 \leq r \leq 1$. □

Remember that this result holds only for the generalized cwatsets in $\mathbf{Z}_2^d$. However, if the pattern for classical cwatsets continues (1/2, 3/4, 7/8,...), and I believe that it does, then this same argument will apply as well.

7 Questions Galore

Classical Cwatsets

• *For which values of k that divide $2^d d!$ do there actually exist cwatsets?* Julie Kerr [3] proved that there is no cwatset of order 15 (which divides $2^5 5!$) in $\mathbf{Z}_2^5$. This problem is a generalization of the problem discussed in section 4.

• *What are some algorithms for constructing cwatsets?* We already have direct sums and cyclic cwatsets. Are there other constructions?

• *What is the supremum for $\rho(C)$ where C is any proper cwatset?*

• *Why does the order of the largest cwatset in $\mathbf{Z}_2^d$ always seem to be of the form $2^d - 2^a$?* This fact is easily seen by examining the list of cwatsets given in Kerr's paper [3]. It is possible that this has to do with the fact that $|C|$ divides $\mathbf{Z}_{2_\phi}^d \times Aut(\mathbf{Z}_2^d) = 2^d(2^d-1)(2^d-2)(2^d-4)\cdots(2^d-2^{d-1})$.

• *Is there a cwatset of order 60 in $\mathbf{Z}_2^6$?* The answer appears to be no. If so, this will prove that the maximum ρ is 3/4 (see section 4).

• *What about a lower bound for the average weight of a cwatset element?*

• *What about the standard deviation of the weights of cwatset elements?* The only standard deviation we have now is for all of $\mathbf{Z}_2^d$ and is $\sigma = \sqrt{d}/2$. A general result, along with the average weight, would give us a better idea of the weight distributions.

Generalized Cwatsets

• *Does there exist an appropriate cwatset such that $\rho(C)$ can be any rational number?* Earlier we showed that we can get $\rho(C)$ to approximate any rational number as close as we want. Can we get the exact fraction?

- *Let $G = \mathbf{Z}_k^d$ and $H = S_d$. If $\mathbf{0} \in C$, will C always be a subgroup?*

References

[1] J. Atkins and G. Sherman, Sets of typical subsamples, Statistics and Probability Letters. **14**, 115-117 (1992).

[2] G. Sherman and M. Wattenberg, Introducing...Cwatsets!, Mathematics Magazine **67**, 109-117 (April 1994).

[3] J. Kerr, Hypergraph Representations and Orders of Cwatsets, Rose-Hulman Technical Report, **MS TR 93-02**, (1993).

[4] A. Forsythe and J. A. Hartigan, Efficiency of confidence intervals generated by repeated subsample calculations, Biometrika **57**, 629-639 (1970).

[5] J. A. Hartigan, Using subsample vakues as typical values, J. Amer. Stat. Assoc. **64**, 1303-1317 (1969).

Appendix A

Another way to prove that $\bar{w} \leq d/2$ is to use dot products. The dot product of cwatset elements is essentially the same as the dot product of vectors: sum the products of the components. For example, suppose $a = 1011$ and $b = 0111$. Then $a \cdot b = 1(0) + 0(1) + 1(1) + 1(1) = 2$.

There are three basic relations that will be used:

$$x \cdot x = \text{weight of } x$$

$$x \cdot y = \text{the number of places where 1's overlap}$$

$$(x + y) \cdot (x + y) = x \cdot x + y \cdot y - 2x \cdot y$$

Now, the total number of ones in a cwatset ($n\bar{w}$)is given by $\sum_x x \cdot x$. However, adding a given element y to the cwatset will not change the total number of ones. So this total is equally well written as $\sum_x (x + y) \cdot (x + y)$. Equating the two sums yields:

$$\begin{aligned} \sum_x x \cdot x &= \sum_x (x + y) \cdot (x + y) \\ &= \sum_x x \cdot x + \sum_x y \cdot y - 2 \sum_x x \cdot y \\ &= \sum_x x \cdot x + n(y \cdot y) - 2 \sum_x x \cdot y \end{aligned}$$

But this implies that

$$n(y \cdot y) = 2 \sum_x x \cdot y$$

Since this is true for any given y, we can sum both sides over all possible y in the cwatset. Doing this gives us

$$\sum_y n(y \cdot y) = 2 \sum_y \sum_x x \cdot y$$

$$n^2 \bar{w} = 2 \sum_y \sum_x x \cdot y = 2 \sum_i w(c_i)^2$$

where $w(c_i)$ is the weight of the i^{th} column. By the Cauchy-Schwarz inequality,

$$d \sum_i w(c_i)^2 \geq [\sum_i w(c_i)]^2 = (n\bar{w})^2$$

So

$$n^2\bar{w} = 2\sum_i w(c_i)^2 \geq \frac{2(n\bar{w})^2}{d}$$

or

$$\bar{w} \leq \frac{d}{2}$$

The reason this proof is included is because the idea of dot products looks as though it might be a fairly powerful technique. The entire proof is based entirely on the fact that the number of ones in a cwatset is 'conserved'. Everything else is just algebra. Hopefully, this technique can be expanded to give more results (possibly including the standard deviation of the weights).

Appendix B

The fact that the weight enumerator of a direct sum is the product of the weight enumerators can be used to decompose cwatsets which are direct sums of other cwatsets. The basic idea is to factor the weight enumerator of the larger cwatset, and then look for smaller cwatsets whose weight enumerators correspond to these factors. For example, consider the cwatset C_3.

0000

0011

1001

0111

0100

1101

It's weight enumerator is $1 + x + 2x^2 + 2x^3 = (1 + x)(1 + 2x^2)$. The cwatset $C_1 = \{0, 1\}$ has the weight enumerator $1 + x$. The cwatset $C_2 = \{000, 101, 011\}$ has the weight enumerator $1 + 2x^2$. And indeed $C_3 = C_1 \oplus C_2$. Well, not exactly. A 'real' direct sum would look like

0000

0101

0011

1000

1101

1011

However, the above cwatset is just $C_3^{(1,2)}$, so the two cwatsets are really the same. Already we see how factoring the weight enumerator can help us find the components of a direct sum even when the cwatset doesn't look 'exactly' like a direct sum.

Caution!

Although this technique is useful, it can also be misleading. First of all, just because a weight enumerator factors, does not mean it is a direct sum. For instance, take the following cwatset, H.

000

100

110

111

011

001

It has the weight enumerator $1 + 2x + 2x^2 + x^3 = (1 + x)(1 + x + x^2)$, but this cwatset is not the direct sum of other cwatsets. There is one other pitfall when looking at weight enumerators. Consider the cwatset, J:

0000

1000

0011

1011

0010

1001

The weight enumerator for J is $(1+x)(1+x+x^2)$ just like for H. However, this cwatset is a direct sum. It is actually $\{0\} \oplus H$. The trick is that the weight enumerator for $\{0\}$ is just 1, so it can be overlooked in the factorization.