On the Statistics of the Number of Fixed-Dimensional Subcubes in a Random Subset of the n-Dimensional Discrete Unit Cube

Svante JANSON, Blair SEIDLER, and Doron ZEILBERGER

Abstract: This paper consists of two independent, but related parts. In the first part we show how to use symbolic computation to derive explicit expressions for the first few moments of the number of implicants that a random Boolean function has, or equivalently the number of fixed-dimensional subcubes contained in a random subset of the n-dimensional unit cube. These explicit expressions suggest, but do not prove, that these random variables are always asymptotically normal. The second part is a full, human-generated proof, of this asymptotic normality.

Motivation

Recall that an *implicant* of a Boolean function in n variables, $f(x_1, \ldots, x_n)$ is a pure disjunction

$$x_{i_1}^{a_1} \wedge x_{i_2}^{a_2} \wedge \ldots \wedge x_{i_r}^{a_r} \quad ,$$

that implies it. Here $1 \leq i_1 < \ldots i_r \leq n$, $a_1, \ldots, a_r \in \{0,1\}$ and $z^1 = z$, and $z^0 = \bar{z}$ (the negation of z).

Fix r and let n vary, We are interested in the statistical distribution of the random variable *number of implicants of length* $n - r$ in a uniformly-at-random Boolean function of n variables. Clearly, when $r = 0$ it is nothing but the good old (fair) binomial distribution with 2^n *fair coin-tosses*, $B(\frac{1}{2}, 2^n)$

Equivalently, for a random subset of the n-dimensional cube, we are interested in statistical distribution of the number of r-dimensional subcubes properly contained in it.

We would like to have **explicit expressions**, in n, for the k^{th} moment of this random variable, for as many as possible r and k. This turns out to be a challenging *symbolic-computational* problem that we will address in the first part of this paper.

In the second part we will prove, using purely human reasoning, that for each r, this distribution is *asymptotically normal*.

Our Random Variables

The *sample space* is the set of subsets of $\{0,1\}^n$, that has cardinality 2^{2^n}.

Let's define our random variables formally.

For a (uniformly-at-) random subset S, of $\{0,1\}^n$, and *fixed* r, define the **random variable**

$X_r(S) :=$ number of r-dimensional subcubes of S.

For example, if $n=3$ and

$$S=\{000,001,010,011,100,111\} \quad ,$$

we have

$$X_0(S)=6 \quad , \quad X_1(S)=6 \quad , \quad X_2(S)=1 \quad , \quad X_3(S)=0 \quad .$$

We would like to get, for as many pairs (k,r) as possible, **explicit** expressions in n, for the k-th moment of X_r, i.e. for

$$f_{kr}(n) := E[{X_r}^k](n) \quad .$$

The Expectation and Variance

The *first* moment, aka *expectation*, aka *mean*, aka *average*, is easy, using **linearity of expectation**.

For any specific subcube C of $\{0,1\}^n$, define the *atomic* random variable, X_C, on subsets, S, of $\{0,1\}^n$ as follows.

$$X_C(S) \,=\, \begin{cases} 1, if \quad C \subset S; \\ 0, \qquad\qquad\qquad \text{otherwise.} \end{cases}$$

Let $\mathcal{C}(n,r)$ be the set of all $\binom{n}{r}2^{n-r}$ r-dimensional subcubes of $\{0,1\}^n$, then, of course

$$X_r(S) \,=\, \sum_{C \in \mathcal{C}(n,r)} X_C(S) \quad .$$

Applying the *expectation* functional, and using the **linearity of expectation**, we get, that the average, let's call it $\mu_r(n)$, is

$$\mu_r(n)=E[X_r] \,=\, E[\sum_{C \in \mathcal{C}(n,r)} X_C] \,=\, \sum_{C \in \mathcal{C}(n,r)} E[X_C] \quad .$$

Now the probability that a random subset of $\{0,1\}^n$ contains an r-dimensional subcube C is $(\frac{1}{2})^{2^r}$, since for each of its vertices, the chance of it belonging to S is $\frac{1}{2}$, and by *independence* the probability that all its 2^r vertices belong to S is indeed $\frac{1}{2^{2^r}}$. The probability that $X_C(S)=0$ is of course $1-(\frac{1}{2})^{2^r}$, hence

$$E[X_C]=1\cdot(\frac{1}{2})^{2^r} \,+\, 0\cdot(1-(\frac{1}{2})^{2^r}) \,=\, (\frac{1}{2})^{2^r} \quad .$$

Going back above we have

$$\mu_r(n) \,=\, \sum_{C \in \mathcal{C}(n,r)} E[X_C] = \sum_{C \in \mathcal{C}(n,r)} \frac{1}{2^{2^r}} = |\mathcal{C}(n,r)|\cdot\frac{1}{2^{2^r}} = \frac{\binom{n}{r}2^{n-r}}{2^{2^r}} \quad .$$

In a beautiful paper, Thanatipanonda [T] derived an explicit expression for the general second moment, for every r-dimensional cube.

Thanatipanoda's General Formula for the Second Moment:

$$E[X_r^2] \;=\; \sum_{i=0}^{r} \frac{n!2^{n-i}}{i!(r-i)!^2(n-2r+i)!2^{2^{r+1}}} \cdot (2^{2^i}-1) \;+\; \frac{[\binom{n}{r}2^{n-r}]^2}{2^{2^{r+1}}} \quad ,$$

from which immediately follows, using $[E(X_r-\mu_r(n))^2] = E[X_r^2] - \mu_r(n)^2$, the following formula.

Thanatipanoda's General Formula for the Variance:

$$Var(X_r) = \sum_{i=0}^{r} \frac{n!2^{n-i}}{i!(r-i)!^2(n-2r+i)!2^{2^{r+1}}} \cdot (2^{2^i}-1) \quad .$$

Note that the variance is a polynomial in $(n, 2^n)$ of degree $2r$ in n and degree 1 in 2^n.

Higher Moments

Edges

Thanatipanonda was unable to get such a general formula for higher moments, but did get $E[X_1^3]$, from which he immediately deduced that the third-moment-about-the-mean of X_1 is

$$E[(X_1-\mu_1(n))^3] \;=\; \frac{3n^3 2^n}{64} \quad .$$

Using the symbolic-computational algorithms to be described in the next section, we managed to derive the following explicit formulas

$$E[(X_1-\mu_1(n))^4] \;=\; \frac{n2^n\left(12\,2^n n^3 + 12\,2^n n^2 + 40n^3 + 3n2^n - 48n^2 + 12n - 16\right)}{1024} \quad .$$

$$E[(X_1-\mu_1(n))^5] \;=\; \frac{5\,2^n n^3\left(6\,2^n n^2 + 3n2^n + 4n^2 - 24n + 8\right)}{1024} \quad .$$

$$E[(X_1-\mu_1(n))^6] \;=$$

$$\frac{n2^n}{32768} \cdot (120\,(2^n)^2\,n^5 + 180n^4\,(2^n)^2 + 1920\,2^n n^5 + 90n^3\,(2^n)^2 - 840n^4 2^n - 1792n^5 + 15n^2\,(2^n)^2$$
$$-360\,2^n n^3 - 5280n^4 - 300\,2^n n^2 + 3840n^3 - 240n2^n + 3840n^2 - 6720n + 4864) \quad .$$

It follows that the **scaled moments about the mean** for the third, fourth, fifth, and sixth moments, converge, as $n \to \infty$, to $0, 3, 0, 15$ respectively, the respective moments of the normal distribution, indicating that the random variable X_1 (the number of edges contained in S) is **probably** asymptotically normal. To fully prove asymptotic normality, of course, we need to prove it for all moments, not just for the first six.

Squares

We only managed to get explicit expressions for the third and fourth moments for X_2.

$$E[(X_2-\mu_2(n))^3] \;=\; \frac{2^n n\,(n-1)\left(9n^4 + 6n^3 + 21n^2 - 16n - 34\right)}{32768} \quad .$$

$$E[(X_2-\mu_2(n))^4] \; = \; \frac{2^n n\,(n-1)}{4194304} \cdot$$
$$(12n^6 2^n + 12\, 2^n n^5 + 520n^6 + 24n^4 2^n - 24n^5 - 12\, 2^n n^3 + 1272n^4 - 9\, 2^n n^2 - 840n^3 - 27n2^n - 5232n^2 - 2768n + 240) \quad .$$

3-dimensional cubes

We only managed to get an explicit expression for the third moment for X_3.

$$E[(X_3-\mu_3(n))^3] \; = \; \frac{2^n n\,(n-1)\,(n-2)\left(14n^6+24n^5+479n^4+2046n^3+6779n^2+15444n-23112\right)}{2415919104}$$

Our Method

Obviously we did not derive these formulas by hand. We had to teach our computer how to find them. It also uses *linearity of expectation*, but with higher moments, things get very complicated. Recall that

$$X_r(S) \; = \sum_{C \in \mathcal{C}(n,r)} X_C(S) \quad .$$

Hence, the k-th moment is

$$E[(X_r)^k] = E[(\sum_{C \in \mathcal{C}(n,r)} X_C(S))^k] =$$

$$\sum_{[C_1,\dots,C_k] \in \mathcal{C}(n,r)^k} E[X_{C_1} X_{C_2} \cdots X_{C_k}]$$

So we sum over all $(\binom{n}{r} 2^{n-r})^k$ members of $\mathcal{C}(n,r)^k$. Since $X_{C_1}(S) X_{C_2}(S) \cdots X_{C_k}(S) = 1$ if **each** of $C_1, C_2, \dots, C_k$ is properly included in S, and 0 otherwise, i.e. if each vertex in $C_1 \cup C_2 \cup \dots C_k$ belongs to S, the contribution due to each such term is

$$E[X_{C_1} X_{C_2} \cdots X_{C_k}] = \frac{1}{2^{|C_1 \cup C_2 \cup \dots C_k|}} \quad .$$

Data Structure

Every r-dimensional subcube of $\{0,1\}^n$ has the form

$$C = \{(x_1,\dots,x_n) \in \{0,1\}^n \,|\, x_{i_1} = \alpha_{i_1}, \dots, x_{i_{n-r}} = \alpha_{i_{n-r}}\} \quad ,$$

for some $1 \le i_1 < i_2 < \dots < i_{n-r} \le n$ and $(\alpha_{i_1}, \dots \alpha_{i_{n-r}}) \in \{0,1\}^{n-r}$. A good way to represent it on a computer is as a row-vector of length n, in the *alphabet* $\{0,1,*\}$, where the entries corresponding to $i_1, i_2, \dots, i_{n-r}$ have $\alpha_{i_1}, \dots . \alpha_{i_{n-r}}$ respectively and the remaining r entries are filled with **wild cards**, denoted by $*$.

For example, if $n=7$ and $r=3$, the 3-dimensional cube

$$\{(x_1,\dots,x_7) \in \{0,1\}^7 \,|\, x_2 = 1, x_4 = 1, x_5 = 0, x_7 = 1\} \quad ,$$

is represented by

$$*1*10*1 \quad .$$

We are trying to find a weighted count of **ordered** k-tuples of r-dimensional subcubes. The natural data structure for these is the set of k by n matrices in the 'alphabet' $\{0,1,*\}$ where every row has exactly r 'wildcards', *.

Let's call this set of matrices, that correspond to $\mathcal{C}(n,r)^k$, $\mathcal{C}(n,k,r)$.

For any **specific**, numeric n, there are 'only' $(2^{n-r}\binom{n}{r})^k$ of these matrices, and for each and every one of them one can find the cardinality of the union of the corresponding subcubes, let's call it v, and add to the running sum $\frac{1}{2^v}$. But we want to do it for **symbolic** n, i.e. for 'all' n. We will soon see how, for each *specific* (numeric) r and k this can be done, *in principle*, but only for relatively small r and k *in practice*. But let's try and push it as far we can. An interesting consequence of our algorithm is the precise degree in n and 2^n of the expression for $E[X_r^k](n)$.

The Kernel

A key object in our approach is the **kernel**. Given a $k \times n$ matrix in the alphabet $\{0,1,*\}$ let's call a column **active** if it contains at least one '$*$'. Note that the matrix has exactly $k \cdot r$ '$*$'s, hence the number of **active columns**, let's call it a, is between r and $k \cdot r$.

[More generally, of we want to find an expression for the *mixed moment* $E[X_{r_1} \cdots X_{r_k}]$ the number of active columns is between $max(r_1, \ldots, r_k)$ and $r_1 + \ldots + r_k$.]

Let $\mathcal{C}_a(n,r,k)$ be the subset of $\mathcal{C}(n,r,k)$ matrices with exactly a active columns. We will call such a matrix in **canonical form** if the active columns are occupied by the a leftmost columns. Let's denote by $\overline{\mathcal{C}}_a(n,r,k)$ the set of such matrices in canonical form. Obviously, there are $\binom{n}{a}$ ways to choose which of the n columns are active and hence

$$Weight(\mathcal{C}(n,k,r)) = \sum_{a=r}^{rk} Weight(\mathcal{C}_a(n,k,r)) = \sum_{a=r}^{rk} \binom{n}{a} Weight(\overline{\mathcal{C}}_a(n,k,r)) \quad .$$

(For any set, S, $Weight(S)$ is the sum of the weights of its members)

Note that this has degree rk in n.

It remains to do a *weighted-count* (where every matrix gets 'credit' $1/2^v$, where v is the cardinality of the union of the corresponding subcubes represented by the k rows, for the set $\overline{\mathcal{C}}_a(n,k,r)$, of matrices in canonical form. Note that there are only **finitely many** choices for the a leftmost columns, i.e. the set of $k \times a$ matrices in the alphabet $\{0,1,*\}$ with the property that every column has at least one '$*$', and every row has exactly r '$*$'s. These can be divided into *equivalence classes* obtained by permuting rows and columns and transposing 0 and 1 in any given column. Once these are sorted into equivalence classes, one needs only examime one representative, and then multiply the weight by the cardinality of the class.

But what about the $n-a$ rightmost columns? Generically they are all distinct, so a good coarse estimate (and upper bound) would be

$$\binom{2^{n-a}}{k} k! \quad .$$

The other extreme is that all the rows of the submatrix consisting of the $n-a$ rightmost columns are identical, and then there are only 2^{n-a} choices to fill them in.

In general, every such member of $\overline{C}_a(n,k,r)$, determines a **set partition** of the set of rows $\{1,\dots,k\}$, if that set-partition has m members $1 \leq m \leq k$, then the number of choices of assigning **different** $0-1$ vectors of length $n-a$ to each of the parts of the set-partition is

$$\binom{2^{n-a}}{m} m! \quad .$$

Now for each a and for each set-partition, we let the computer generate the **finite** set of $k \times a$ matrices in the alphabet $\{0,1,*\}$. Each of the members of the set partition has its own submatrix, and we ask our computer to kindly find the number of vertices in the corresponding union of subcubes corresponding to each member of the examined set partition. Since they are disjoint, we add them up, getting v for that particular pair (matrix, set-partition), giving credit $1/2^v$.

Implementation

All this is implemented in the Maple package `SMCboole.txt`, available from:

`https://sites.math.rutgers.edu/~zeilberg/tokhniot/SMCboole.txt` .

In particular '`Moms(A,n);`', for any list of non-negative integers $A=[r_1,\dots r_k]$ gives you the mixed moment $E[X_{r_1}\cdots X_{r_k}]$. For example, to get the third moment of the number of edges (i.e. 1-dimensional subcubes) type

`Moms([1,1,1],n);` ,

getting, very fast:

$$\frac{2^n n^2 \left((2^n)^2 n + 12\, 2^n n + 6\, 2^n + 24n \right)}{512} \quad .$$

To get the third moment of the number of squares (i.e. 2-dimensional subcubes), type

Moms([2,2,2],n); ,

getting

$$\frac{2^n n\,(n-1)}{2097152} \cdot$$

$$\left((2^n)^2 n^4 - 2\,(2^n)^2 n^3 + 48\, 2^n n^4 + (2^n)^2 n^2 + 576 n^4 + 24\, 2^n n^2 + 384 n^3 - 72\, 2^n n + 1344 n^2 - 1024 n - 2176\right) \quad .$$

The third moment of the number of 3-dimensional cubes takes a bit longer, and we were unable to compute the fourth moment of the number of 3-dimensional cubes, it took too much time and too much space.

More informative for *statistical purposes* are the **moments about the mean**, $E[(X_r - \mu_r(n))^k]$, that Maple easily derives, using *linearity of expectation* from the pure moments. The function call for this is

`MOMrk(r,k,n);` ,

where `r` and `k` are *numeric* but n is a **symbol** denoting the dimension of the ambient cube. To get the explicit expression given above for the third-through-the six moments of the number of edges, the third and fourth for the number of squares, and the third moment for the number of 3-dimensional subcube (all *about the mean*) we typed:

`MOMrk(1,3,n);` , `MOMrk(1,4,n);` , `MOMrk(1,5,n);` , `MOMrk(1,6,n);` ,

`MOMrk(2,3,n);` , `MOMrk(2,4,n);` ,

`MOMrk(3,3,n);` ,

respectively. To our chagrin, '`MOMrk(3,4,n);`' took too long.

Consequence of the algorithm: The k-th moment of X_r is a bivariate polynomial in $(n, 2^n)$ of degree $k\,r$ in n and degree k in 2^n .

This raises the *theoretical* possibility (in God's computer) of finding these expressions by **pure brute force**. The generic polynomial in $(n, 2^n)$ of degree kr in n and degree k in 2^n has $(1 + kr)(1 + k)$ 'degrees of freedom'. So using *undetermined coefficients* we need to generate a table of $E[X_r^k](n)$ for $1 \leq n \leq (1 + kr) \cdot (1 + k)$. After gathering the data, we use linear algebra to solve a system of $(1 + kr) \cdot (1 + k)$ equations with that many unknowns. For each specific $n = n_1$ there are 'only' $2^{2^{n_1}}$ subsets, and for each of them we can ask how many r-dimensional subcubes do they contain, raise it to the k-th power and take the average. Alas 2^{2^5} is already big enough, so only God's computer, with practically infinite time and space, can carry this brute force approach.

References

[T] Thotsaporn Aek Thanatipanonda, *Reviews of Symbolic Moment Calculus*, arXiv:2003.11749. `https://arxiv.org/abs/2003.11749`

Svante Janson, Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden. `svante.janson@math.uu.se`

Blair Seidler and Doron Zeilberger, Department of Mathematics, Rutgers University (New Brunswick),

Hill Center-Busch Campus, 110 Frelinghuysen Rd., Piscataway, NJ 08854-8019, USA.
Email: `blair.seidler@math.rutgers.edu` , `DoronZeil@gmail.com` .

Written: **Feb. 10, 2023**.