Probability in the Engineering and Informational Sciences, **18**, 2004, 161–183. Printed in the U.S.A.

A GENERAL APPROACH TO COMPUTE THE PROBABILITIES OF UNRESOLVED CLONES IN RANDOM POOLING DESIGNS*

F. K. Hwang and Y. C. Liu

Department of Applied Mathematics
National Chiao Tung University
Hsinchu 30050 Taiwan, Republic of China
E-mail: fhwang@math.nctu.edu.tw;
u8722518@math.nctu.edu.tw

In this paper, we develop a general approach to compute the probabilities of unresolved clones in random pooling designs. This unified and systematic approach gives better insight for handling the dependency issue among the columns and among the rows. Consequently, we identify some faster computation formulas for four random pooling designs proposed in the literature, and we derive some probability distribution functions of the number of unresolved clones that were not available before.

1. INTRODUCTION

A *pooling design* is a (binary) incidence matrix where each column represents a clone and each row represents a pool. A 1-entry in cell (i, j) signifies that clone j is contained in pool i. A clone represents a short DNA fragment. It can be a *positive* if it contains a specific DNA sequence as a subsequence, or a *negative* if otherwise. All clones contained in a pool are tested together as a group. The *outcome* is positive if and only if the pool contains a positive (otherwise, the outcome is negative). The goal of a pooling design is to identify all positives with as few pools as possible. Note that all pools in the matrix can be tested simultaneously.

Random pooling designs have been proposed [1–5] for their wide applicability. However, they do not guarantee to identify all clones. Suppose that there are d

*Partially supported by Republic of China, National Science Council grant NSC 90-2115-M-009-027.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

positives among n clones. Let $\bar{N}$ denote the numbers of *unresolved negatives* and $\bar{P}$ the number of *unresolved positives*. Then, it is important to estimate $\bar{P}$ and $\bar{N}$ for evaluating a random pooling design.

Four types of random design have been studied in the literature. Consider a $t \times n$ binary matrix M:

1. *Random incidence design* (RID). Each cell in M has probability p of being one.
2. *Random k-set design* (RkSD). Each column is a random k-set of the set $[t] = \{1, \dots, t\}$; that is, there are exactly k 1-entries in each column, with the locations of these 1's being equally likely to be any of the $\binom{t}{k}$ possibilities.
3. *Random distinct k-set design* (RDkSD). RkSD with all columns being distinct.
4. *Random r-size design* (RrSD). Each row is a random r-set of the set $[n] = \{1, \dots, n\}$.

RID was first proposed by Erdös and Rényi [3] in search problems. Balding, Bruno, Knill, and Torney [1] proposed RkSD and showed it to have much smaller $\bar{P}$ and $\bar{N}$ (i.e., it identified many more positive clones); hence, it is much more powerful than RID. Both [1] and [2] alluded to bounding of intersections (pools in which two columns coincide) of two columns to avoid heavy similarity without giving any analysis. Hwang [4] carried out the idea by studying the RDkSD in which the sampling is without replacement to avoid some apparent inefficiency. Another motivation to study RDkSD is that some other pooling designs [6,7] are special cases of RDkSD. Hwang also proposed RrSD to compare its row structure with the column structure of RkSD. Further, RrSD is suitable for the situation when the size of a pool is restricted. Note that one can also propose random distinct r-size design (RDrSD). However, since Hwang and Liu [5] found that the performances of RkSD and RDkSD are about the same, and the performance of RrSD is much worse, there is not much motivation to study RDrSD.

The basic probabilities to be computed for these four models are $P(\bar{N} = u)$ and $P(\bar{P} = v)$. In particular, we are interested in $P(\bar{N} = 0)$ and $P(\bar{P} = 0)$, the cases that all negatives and all positives, respectively, are identified. From $P(\bar{N} = u)$ and $P(\bar{P} = v)$ we also obtain $E(\bar{N})$ and $E(\bar{P})$, which in turn yield the following

$$P^- \equiv \frac{E(\bar{N})}{n-d}, \text{ the probability that a random negative clone is unresolved and}$$

$$P^+ \equiv \frac{E(\bar{P})}{d}, \text{ the probability that a random positive clone is unresolved.}$$

However, since $E(\bar{N})$ and $E(\bar{P})$ are usually quite messy, it is often easier to argue for P^- and P^+ directly.

We will also look into the problem of choosing p, k, and r to minimize these probabilities. Due to the messiness of the probability function, only limited results have been obtained.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

2. A GENERAL APPROACH TO COMPUTE THE PROBABILITIES OF UNRESOLVED CLONES

Let M be a $t \times n$ matrix and D a set of d positives. We say that the rows (columns) are i.d. if the distribution of the number of 1-entries are identical for all rows (columns). Furthermore, if the distribution is independent between the rows (columns), then we say the rows (columns) are i.i.d. In the random designs we study here, the rows and the columns are always i.d.

We say a row and a column *intersect* if the intersection cell contains a 1-entry. Partition the rows of M into three parts X, Y, Z according to whether a pool intersects D at least twice, exactly once, or not. Define $f(z) = P(|Z| = z)$.

Theorem 2.1: *Suppose that the rows are i.d. Then,*

$$f(z) = \binom{t}{z} \sum_{i=z}^{t} (-1)^{i-z} \binom{t-z}{i-z} \times P(i \textit{ specified rows including } Z \textit{ not intersecting } D).$$

Proof: $f(z)$ is computed by an inclusion–exclusion formula. ■

Corollary 2.2: *Suppose that the rows are i.d. and the columns are i.i.d. Then,*

$$f(z) = \binom{t}{z} \sum_{i=z}^{t} (-1)^{i-z} \binom{t-z}{i-z} \times [P(a \textit{ column not intersecting the } i \textit{ rows including } Z)]^{d}.$$

Proof: The event "the i rows including Z not intersecting D" can also be expressed as "d columns not intersecting the i rows including Z." ■

When the rows are i.i.d., then $f(z)$ is much simpler.

Lemma 2.3: *Suppose that the rows are i.i.d. Then,*

$$f(z) = \binom{t}{z} [P(a \textit{ row not intersecting } D)]^{z} [1 - P(a \textit{ row not intersecting } D)]^{t-z}.$$

Next, we give $P(\bar{N} = u)$ and then $P(\bar{P} = v)$.

Theorem 2.4: *Suppose that the columns and rows are both i.d. Then,*

$$P(\bar{N} = u) = \sum_{z=0}^{t} f(z) \binom{n-d}{u} P(\textit{exactly } u \textit{ negative columns not in } Z).$$

Proof: A negative column not in Z is an unresolved negative. ■

Either the column i.i.d. or the row i.i.d. will bring some simplification.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

Corollary 2.5: *Suppose that the rows are i.d. and the columns are i.i.d. Then,*

$$P(\bar{N}=u)=\sum_{z=0}^{t} f(z)\binom{n-d}{u}[P(a\ negative\ column\ not\ in\ Z)]^{u}$$
$$\times [P(a\ negative\ column\ in\ Z)]^{n-d-u}.$$

Corollary 2.6: *Suppose that the columns are i.d. and the rows are i.i.d. Then,*

$$P(\bar{N}=u)=\sum_{z=0}^{t} f(z)\binom{n-d}{u}\sum_{j=u}^{n-d}(-1)^{j-u}\binom{n-d-u}{j-u}$$
$$\times \begin{bmatrix} P(a\ pool\ in\ Z\ does\ not\ contain\ any\ of\ the \\ j\ specified\ negative\ clones\ including\ \bar{N}) \end{bmatrix}^{z}.$$

Let $f(z,y)$ denote $P(|Z|=z,|Y|=y)$ and q_S the number of rows in S not containing any unresolved negative.

Theorem 2.7: *Suppose that the columns and rows are both i.d. Then,*

$$P(\bar{p}=v)=\sum_{z=0}^{t}\sum_{y=0}^{t-z} f(z,y)\sum_{u=0}^{n-d} P(\bar{N}=u|z)\sum_{q_Y=0}^{y} P(q_Y|y,u)P(\bar{P}=v|q_Y),$$

where

$$P(\bar{P}=v|q_S)=\binom{d}{v}d^{-q_S}\sum_{l=0}^{d-v}(-1)^{l}\binom{d-v}{d-v-l}(d-v-l)^{q_S}.$$

Proof: Note that $P(\bar{P}=v|q_Y)$ is the probability that v positives do not appear in the q_Y rows, hence unresolved. $P(\bar{P}=v|q_Y)$ can also be viewed as the probability of getting v empty holes in rolling q_Y balls into d holes. ■

Although, in principle, the computation of y and q can be combined into one step, such a computation would be difficult to carry out unless the rows are independent. Here, we give the i.i.d. version.

Corollary 2.8: *Suppose that the columns are i.d. and the rows are i.i.d. Then,*

$$P(\bar{P}=v)=\sum_{z=0}^{t} f(z)\sum_{u=0}^{n-d} P(\bar{N}=u|z)\sum_{q_{X\cup Y}=0}^{t-z}\binom{t-z}{q_{X\cup Y}}$$
$$\times [P(E)]^{q_{X\cup Y}}[1-P(E)]^{t-z-q_{X\cup Y}}P(v|q_{X\cup Y}),$$

where E is the event that a row in $X\cup Y$ contains a single positive but no unresolved negative.

We next discuss the computation of P^{-} and P^{+}. For convenience, in computing P^{-}, the negative whose resolvability is in concern will be denoted by C. In computing P^{+}, D_1 is the positive in concern. We first give a general formula for computing P^{-}.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

Theorem 2.9: *Suppose that the rows are i.d. Then,* $P^-(C) = \sum_{z=0}^{t} f(z) P(Z \textit{ does not contain } C)$.

Corollary 2.10: P^- *is independent of* n *if and only if* $f(z)$ *is independent of* n.

Corollary 2.11: P^- *is independent of* n *if the columns are independent.*

Proof: $f(z)$ is determined by the columns of D, hence independent of n if the columns are independent. ■

Corollary 2.12: *Suppose that the rows are i.i.d. Then,*

$$P^-(C) = \sum_{z=0}^{t} f(z)[P(a \textit{ pool of } Z \textit{ does not contain } C)]^z.$$

However, we can do better by combining the computation of the probabilities of z and of the property.

Theorem 2.13: *Suppose that the rows are i.i.d. Then,*

$$P^-(C) = [1 - P(a \textit{ pool contains } C, \textit{ but none of } D)]^t.$$

Even without the row i.d., we can interpret Theorem 2.9 in a way such that there is no need to compute $f(z)$. Let the *weight* of a column be the number of its 1-entries.

Theorem 2.14: *Suppose that* C *has weight* k. *Then,*

$$P^-(C) = \sum_{i=0}^{k} (-1)^i \binom{k}{i} P(\textit{the } i \textit{ specified rows each not intersecting } D).$$

Proof: The i specified rows are in the k rows contained in C. ■

Corollary 2.15: *Suppose that the columns are i.i.d. Then,*

$$P^-(C) = \sum_{i=0}^{k} (-1)^i \binom{k}{i} [P(\textit{the } i \textit{ appearances of } C \textit{ does not intersect } D_1)]^d.$$

To compute P^+, let Y_1 be the subset of Y containing D_1 but no other D_j. Define

$$f(z, y_1) = P(|Z| = z, |Y_1| = y_1).$$

Theorem 2.16: *Suppose that the rows are i.d. Then,*

$$P^+(D_1) = \sum_{z=0}^{t} \sum_{y_1=0}^{t-z} f(z, y_1) \sum_{i=0}^{y_1} (-1)^i \binom{y_1}{i} \times P\begin{pmatrix} \textit{all negatives either appearing in } Z \textit{ or not} \\ \textit{appearing in the } i \textit{ specified rows of } Y_1 \end{pmatrix}.$$

Proof: The last sum gives the probability that no row in Y_1 satisfies the condition that every negative either appears in Z (hence resolved) or does not appear in the row of Y_1 (hence not obstructing the identification of D_1). Note that the condition characterizes the identification of D_1. ■

COROLLARY 2.17: *Suppose that the rows are i.d. and the columns are i.i.d. Then,*

$$P^+(D_1) = \sum_{z=0}^{t}\sum_{y_1=0}^{t-z} f(z,y_1)\sum_{i=0}^{y_1}(-1)^i\binom{y_1}{i}$$
$$\times \left[P(a \text{ negative appears in } Z) + P\begin{pmatrix} a \text{ negative does not appear in } Z, \\ \text{nor in the specified rows of } Y_1 \end{pmatrix}\right]^{n-d}$$

With the row i.i.d., we obtain a different set of formulas.

THEOREM 2.18: *Suppose that the rows are i.i.d. Then,*

$$P^+(D_1) = \sum_{z=0}^{t} f(z)\sum_{u=0}^{n-d} P(\bar{N} = u|z)$$
$$\times \left[1 - P\begin{pmatrix} a \text{ row in } X \cup Y \text{ contains } D_1 \text{ but no other } D_i, \\ \text{nor any of the } u \text{ unresolved negatives} \end{pmatrix}\right]^{t-z}.$$

PROOF: The [ ] term is the probability that D_1 is not identified by any of the $t - z$ rows, hence unresolved. ■

THEOREM 2.19: *Suppose that the columns are i.d. and the rows are i.i.d. Then,*

$$P^+(D_1) = \sum_{u=0}^{n-d}\binom{n-d}{u}\sum_{j=u}^{n-d}(-1)^{j-u}\binom{n-d-u}{j-u}$$
$$\times \left[P(a \text{ pool is positive and does not identify } D_1 \text{ given } \bar{N} = u) + P\begin{pmatrix} a \text{ pool is negative and does not contain a} \\ \text{specified set of } j \text{ negatives including } \bar{N} \end{pmatrix}\right]^{t}.$$

PROOF:

$$P^+(D_1) = \sum_{u=0}^{n-d} P(\bar{N} = u)P(P^+(D_1)|\bar{N} = u)$$
$$= \sum_{u=0}^{n-d}\binom{n-d}{u}\sum_{j=u}^{n-d}(-1)^{j-u}\binom{n-d-u}{j-u}$$
$$\times P(\text{besides } \bar{N}, j - u \text{ additional negative clones not in } Z)$$
$$\times P(P^+(D_1)|\bar{N} = u).$$

The second equality is true by Corollary 2.6.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

Define the following:

A: the event that besides $\overline{N}$, exactly $j - u$ additional clones not in Z

B: the event D_1 not identified given $\overline{N} = u$

Since B depends on D and $\overline{N}$, and A depends on $j - u$ clones not in $D \cup \overline{N}$, A and B are independent events. Hence,

$$\begin{aligned} P(A)P(B) &= P(AB) \\ &= \left[P\left(\begin{array}{l} \textit{a pool does not identify } D_1 \textit{ given } \overline{N} = u. \\ \textit{If the pool is negative, then a specified} \\ \textit{set of } j \textit{ negatives including } \overline{N} \textit{ is not in it} \end{array} \right) \right]^t \\ &= \Bigg[P(\textit{a pool is positive and does not identify } D_1, \textit{ given } \overline{N} = u) \\ &\qquad + P\left(\begin{array}{l} \textit{a pool is negative and does not contain a} \\ \textit{specified set of } j \textit{ negatives including } \overline{N} \end{array} \right) \Bigg]^t. \end{aligned}$$ ■

Theorem 2.19 can also be obtained from Theorem 2.18 by replacing $P(\overline{N} = u|z)$ with the terms in Corollary 2.6 and summing over z. The proof we gave here is more insightful.

3. RANDOM INCIDENCE DESIGN

Let M be a $t \times n$ RID. Note that both rows and columns are i.i.d. Using the row i.i.d., the following is easily obtained:

LEMMA 3.1:

$$f(z) = \binom{t}{z} (1-p)^{dz} [1 - (1-p)^d]^{t-z}.$$

Hwang [4] gave the following theorem.

THEOREM 3.2:

$$\begin{aligned} P(\overline{N} = u) = \sum_{z=0}^{t} & \binom{t}{z} (1-p)^{dz} [1 - (1-p)^d]^{t-z} \\ & \times \binom{n-d}{u} (1-p)^{zu} [1 - (1-p)^z]^{n-d-u}. \end{aligned}$$

PROOF: The proof follows immediately from Corollary 2.5 and Lemma 3.1. ■

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

The special case $u = 0$ was first given by Balding et al. [1].

COROLLARY 3.3:

$$P(\bar{N} = 0) = \sum_{z=0}^{t} \binom{t}{z} (1-p)^{dz} [1-(1-p)^d]^{t-z} [1-(1-p)^z]^{n-d}.$$

COROLLARY 3.4: $E(\bar{N}) = (n-d)[1-p(1-p)^d]^t$.

PROOF:

$$\begin{aligned} E(\bar{N}) &= \sum_{z=0}^{t} \binom{t}{z} (1-p)^{dz} [1-(1-p)^d]^{t-z} \\ &\quad \times \sum_{u=0}^{n-d} u \binom{n-d}{u} (1-p)^{zu} [1-(1-p)^z]^{n-d-u} \\ &= \sum_{z=0}^{t} \binom{t}{z} (1-p)^{dz} [1-(1-p)^d]^{t-z} (n-d)(1-p)^z \\ &= (n-d)\left[\sum_{z=0}^{t} \binom{t}{z} (1-p)^{(d+1)z} [1-(1-p)^d]^{t-z}\right] \\ &= (n-d)[(1-p)^{d+1} + 1 - (1-p)^d]^t \\ &= (n-d)[1-p(1-p)^d]^t. \end{aligned}$$

■

COROLLARY 3.5: $P^- = [1-p(1-p)^d]^t$.

Note that Corollary 3.5 can also be argued directly from Theorem 2.13 by noting that $p(1-p)^d$ is the probability that a row contains C but none of D. Then, Corollary 3.5 can be obtained by multiplying by $(n-d)$. We did it the hard way just for demonstration purposes.

Let p_*^- minimize P^- (or $E(\bar{N})$). Balding et al. [1] gave the following:

THEOREM 3.6: $p_*^- = (d+1)^{-1}$.

PROOF: Clearly, to minimize P^- is to maximize $p(1-p)^d$. Set

$$\frac{d}{dp} p(1-p)^d = (1-p)^d - pd(1-p)^{d-1} = 0.$$

We obtain $p_*^- = (d+1)^{-1}$. ■

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

Let p_0^- minimize $P(\bar{N}=0)$. No analytic solution of p_0^- is known.

The corresponding probabilities of unresolved positives are considerably messier.

THEOREM 3.7:

$$\begin{aligned} P(\bar{P}=v) = {} & \sum_{z=0}^{t} \binom{t}{z} (1-p)^{dz} [1-(1-p)^d]^{t-z} \\ & \times \binom{n-d}{u} \sum_{u=0}^{n-d} (1-p)^{zu} [1-(1-p)^z]^{n-d-u} \\ & \times \sum_{q=0}^{t-z} \binom{t-z}{q} \left[\frac{dp(1-p)^{d-1+u}}{1-(1-p)^d} \right]^q \left[1 - \frac{dp(1-p)^{d-1+u}}{1-(1-p)^d} \right]^{t-z-q} \\ & \times \binom{d}{v} d^{-q} \sum_{l=0}^{d-v} (-1)^l \binom{d-v}{d-v-l} (d-v-l)^q. \end{aligned}$$

PROOF: This is proved by Corollary 2.8. ■

Because $P(\bar{P}=v)$ is unwieldy to maneuver, it is desirable to derive P^+ and $E(\bar{P})$ independently. We give several such derivations and compare their terms' complexities. First, a lemma is needed.

LEMMA 3.8:

$$f(z, y_1) = \binom{t}{z, y_1} (1-p)^{dz} [p(1-p)^{d-1}]^{y_1} [1-(1-p)^{d-1}]^{t-z-y_1}.$$

PROOF: A pool is not in $Z \cup Y_1$ if and only if it contains a positive other than D_1. ■

We can use the column i.i.d. to compute P^+.

THEOREM 3.9:

$$\begin{aligned} P^+ = {} & \sum_{z=0}^{t} \sum_{y_1=0}^{t-z} \binom{t}{z, y_1} (1-p)^{dz} [p(1-p)^{d-1}]^{y_1} [1-(1-p)^{d-1}]^{t-z-y_1} \\ & \times \sum_{i=0}^{y_1} (-1)^i \binom{y_1}{i} [1-(1-p)^z + (1-p)^{z+i}]^{n-d}. \end{aligned}$$

PROOF: $1-(1-p)^z$ is the probability that a negative appears in Z, and $(1-p)^{z+i}$ is the probability that a negative does not appear in Z or in the i specified rows of Y_1. Theorem 3.9 follows immediately from Corollary 2.17. ■

Note that P^+ in Theorem 3.9 can be computed in $O(t^3)$ time.

Alternatively, we can use the row i.i.d. formula in Corollary 2.6 (after summing over z).

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

THEOREM 3.10:

$$P^+ = \sum_{u=0}^{n-d} \binom{n-d}{u} \sum_{j=u}^{n-d} (-1)^{j-u} \binom{n-d-u}{j-u} \times [1-(1-p)^d - p(1-p)^{d-1+u} + (1-p)^{d+j}]^t.$$

PROOF: $1-(1-p)^d$ is the probability that a pool contains a positive; hence, it is positive. In a positive pool, D_1 is identified if and only if it is the only positive in the pool and no unresolved negative is in the pool. The probability of this is $p(1-p)^{d-1+u}$, given there are u negative pools. Therefore, $1-(1-p)^d - p(1-p)^{d-1+u}$ is the probability that a pool is positive but not identifying D_1 given $\bar{N} = u$.

On the other hand, $(1-p)^d$ is the probability that a pool is negative and $(1-p)^j$ is the probability that it does not contain the j specified negatives including $\bar{N}$. Hence, $(1-p)^{d+j}$ is the probability that both events happen. Theorem 3.10 follows immediately from Theorem 2.19. ■

Note that P^+ in Theorem 3.10 can be computed in $O(n^2)$ time.

Finally, we can also use the other row i.i.d. formula.

THEOREM 3.11:

$$P^+ = \sum_{z=0}^{t} \binom{t}{z} (1-p)^{dz} [1-(1-p)^d]^{t-z} \times \sum_{u=0}^{n-d} \binom{n-d}{u} (1-p)^{zu} [1-(1-p)^z]^{n-d-u} \times \left[1 - \frac{p(1-p)^{d-1+u}}{1-(1-p)^d}\right]^{t-z}.$$

PROOF: $p(1-p)^{d-1+u}$ is the unconditional probability that a pool contains D_1 but no other D_i nor any unresolved negative. Its division by $1-(1-p)^d$ given the same probability conditional on the pool is positive (in $X \cup Y$). Theorem 3.11 follows immediately from Theorem 2.18. ■

P^+ in Theorem 3.11 can be computed in $O(tn)$ time. Since t is usually much smaller than n, Theorem 3.11 seems to be an improvement over Theorem 3.10 with respect to computation. Note that Theorem 3.9 uses the column independence with time complexity a function of t, Theorem 3.10 uses the row independence with time complexity a function of n, and Theorem 3.11 uses both column and row independence with time complexity a function of both t and n.

COROLLARY 3.12: $E(\bar{P}) = dP^+$

No analytic solution has been given to minimize either P^+ or $P(\bar{P} = 0)$.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

4. RANDOM *k*-SET DESIGN

The columns in RkSD are i.i.d., but the rows are only i.d.

LEMMA 4.1:

$$f(z) = \binom{t}{z} \sum_{i=z}^{t} (-1)^{i-z} \binom{t-z}{i-z} \left[\frac{\binom{t-i}{k}}{\binom{t}{k}} \right]^d .$$

PROOF: Since the rows are not independent, the inclusion–exclusion formula is used to compute the exact probability of z. ■

THEOREM 4.2:

$$P(\bar{N} = u) = \sum_{z=0}^{t} \binom{t}{z} \sum_{i=z}^{t} (-1)^{i-z} \binom{t-z}{i-z} \left[\frac{\binom{t-i}{k}}{\binom{t}{k}} \right]^d \times \binom{n-d}{u} \left[\frac{\binom{t-z}{k}}{\binom{t}{k}} \right]^u \left[1 - \frac{\binom{t-z}{k}}{\binom{t}{k}} \right]^{n-d-u} .$$

PROOF: The probability that a negative does not appear in a row of Z is $\binom{t-z}{k} \Big/ \binom{t}{k}$. Theorem 4.2 now follows immediately from Corollary 2.4. ■

It is easier to argue for P^- independently than from $P(\bar{N} = u)$.

THEOREM 4.3:

$$P^- = \sum_{i=0}^{k} (-1)^i \binom{k}{i} \left[\frac{\binom{t-i}{k}}{\binom{t}{k}} \right]^d .$$

PROOF: The probability that a positive does not appear in i of the k appearances of C is $\binom{t-i}{k} \Big/ \binom{t}{k}$. Theorem 4.3 follows immediately from Corollary 2.15. ■

COROLLARY 4.4: $E(\bar{N}) = (n-d)P^-$.

Let k_*^- minimize P^-. Our formula for P^- is very similar to that Macula [8] gave for the probability of a positive being unresolved under the representative decoding [6]. Hence, we imitate the approximation he gave:

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

$$\begin{aligned}
P^- &\sim \sum_{i=0}^{k} (-1)^i \binom{k}{i} \left(1 - \frac{i}{t}\right)^{kd} \\
&\sim \sum_{i=0}^{k} (-1)^i \binom{k}{i} e^{(-kd/t)i} \\
&= (1 - e^{(-kd/t)})^k.
\end{aligned}$$

Then, $k' = (t \ln 2)/d$ minimizes $(1 - e^{-kd/t})^k$.

To compute $P(\bar{P} = v)$, we need $f(z, y)$. Let $\Pi(y, d)$ denote the set of partitions $\pi = y_1, \ldots, y_d$ of $y = \sum_{j=1}^{d} y_j$ distinct objects into d distinct parts with $0 \le y_j \le k$.

To compute P^+, we need $f(z, y_1)$.

Lemma 4.7:

$$f(z, y_1) = \binom{t}{z, y_1} \frac{\binom{t - z - y_1}{k - y_1}}{\binom{t}{k}} \sum_{h=0}^{t-z-y_1} (-1)^h \binom{t - z - y_1}{h} \left[\frac{\binom{t - z - y_1 - h}{k}}{\binom{t}{k}} \right]^{d-1}.$$

Proof: By definitions of z and y_1, each of the remaining $t - z - y_1$ pools must contain a D_i, $i \neq 1$. The last sum in Lemma 4.7 gives this probability using the inclusion–exclusion formula, where $\binom{t-z-y_1-h}{k} \big/ \binom{t}{k}$ is the probability that D_i does not appear in a specified set of $z + y_1 + h$ pools (including the pools in $Z \cup Y_1$). Finally, D_1 must appear in the y_1 rows of Y_1. Its other $k - y_1$ appearances must not be in $Z \cup Y_1$. ■

Theorem 4.8:

$$\begin{aligned}
P^+(D_1) = \sum_{z=0}^{t} \sum_{y_1=0}^{k} \binom{t}{z, y_1} \frac{\binom{t - z - y_1}{k - y_1}}{\binom{t}{k}} \\
\times \sum_{h=0}^{t-z-y_1} (-1)^h \binom{t - z - y_1}{h} \left[\frac{\binom{t - z - y_1 - h}{k}}{\binom{t}{k}} \right]^{d-1} \\
\times \sum_{i=0}^{y_1} (-1)^i \binom{y_1}{i} \left[\frac{\binom{t}{k} - \binom{t - z}{k} + \binom{t - z - i}{k}}{\binom{t}{k}} \right]^{n-d}.
\end{aligned}$$

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

Proof: $\left[\binom{t}{k} - \binom{t-z}{k}\right] \Big/ \binom{t}{k}$ is the probability that a negative appears in Z. $\binom{t-z-i}{k} \Big/ \binom{t}{k}$ is the probability that a negative does not appear in z or in the i specified pools of Y_1. Theorem 4.8 follows immediately from Corollary 2.17. ■

Note that P^+ in Theorem 4.8 can be computed in $O(t^2k^2)$ time.

Corollary 4.9: $E(\bar{P}) = dP^+$.

No analytic solution has been given to minimize either P^+ or $P(\bar{P} = 0)$.

5. RANDOM r-SIZE DESIGN

The rows of RrSD are i.i.d., but the columns are only i.d.

Lemma 5.1:

$$f(z) = \binom{t}{z}\left[\frac{\binom{n-d}{r}}{\binom{n}{r}}\right]^{z}\left[1 - \frac{\binom{n-d}{r}}{\binom{n}{r}}\right]^{t-z}.$$

Proof: $\binom{n-d}{r} \Big/ \binom{n}{r}$ is the probability that a pool does not contain any positive; hence, it is in Z. ■

Theorem 5.2:

$$P(\bar{N} = u) = \sum_{z=0}^{t}\binom{t}{z}\left[\frac{\binom{n-d}{r}}{\binom{n}{r}}\right]^{z}\left[1 - \frac{\binom{n-d}{r}}{\binom{n}{r}}\right]^{t-z} \times \binom{n-d}{u}\sum_{j=u}^{n-d}(-1)^{j-u}\binom{n-d}{j-u}\left[\frac{\binom{n-d-j}{r}}{\binom{n-d}{r}}\right]^{z}.$$

Proof: $\binom{n-d-j}{r} \Big/ \binom{n-d}{r}$ is the probability that a pool in Z does not contain any of the j specified negatives, including the given u ones. Theorem 5.2 now follows immediately from Corollary 2.6 and Lemma 5.1. ■

It is simpler to derive P^- directly.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

THEOREM 5.3:

$$P^- = \left[1 - \frac{\binom{n-d-1}{r-1}}{\binom{n}{r}}\right]^t.$$

PROOF: A pool contains C, but none of D must take its other $r-1$ clones from the other $n-d-1$ negatives. Theorem 5.3 now follows immediately from Theorem 2.13. ■

Let r_*^- minimize P^-. Lin (private communication) observed the following:

THEOREM 5.4: $r_*^- \in \{\lceil r^* \rceil, \lfloor r^* \rfloor\}$ *where* $r^* = (n-d)/(d+1)$.

PROOF: Clearly, minimizing P^- is the same as maximizing $\binom{n-d-1}{r-1}\Big/\binom{n}{r} \equiv g(r)$.

$$\frac{g(r+1)}{g(r)} = \frac{\binom{n-d-1}{r}\binom{n}{r}}{\binom{n-d-1}{r-1}\binom{n}{r+1}} = \frac{(n-d-r)(r+1)}{r(n-r)} = \left(1 - \frac{d}{n-r}\right)\left(1 + \frac{1}{r}\right).$$

When r increases, both factors decrease. Hence, the ratio decreases in r, and maximum $g(r)$ is obtained at the two integers that flank the r^* satisfying $g(r+1)/g(r) = 1$; that is, $r^* = (n-d)/(d+1)$. ■

Note that r^* divided by the number of negatives yields P_*^- in RID.

For the unresolved positive, we have the following theorem.

THEOREM 5.5:

$$P(\bar{P} = v) = \sum_{z=0}^{t} \binom{t}{z} \left[\frac{\binom{n-d}{r}}{\binom{n}{r}}\right]^z \left[1 - \frac{\binom{n-d}{r}}{\binom{n}{r}}\right]^{t-z} \sum_{u=0}^{n-d} \binom{n-d}{u} \sum_{j=u}^{n-d} (-1)^{j-u}$$

$$\times \binom{n-d-u}{j-u} \left[\frac{\binom{n-d-j}{r}}{\binom{n-d-u}{r}}\right]^z \sum_{q=0}^{t-z} \binom{t}{q} \left[\frac{d\binom{n-d-u}{r-1}}{\binom{n}{r} - \binom{n-d}{r}}\right]^q$$

$$\times \left[1 - \frac{d\binom{n-d-u}{r-1}}{\binom{n}{r} - \binom{n-d}{r}}\right]^{t-z-q} \binom{d}{v} d^{-q} \sum_{l=0}^{d-v} (-1)^l \binom{d-v}{d-v-l} (d-v-l)^q.$$

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

Proof: Theorem 5.5 follows from Corollary 2.8. We will only comment on the term $P(E)$ (defined as in Corollary 2.8), as the other terms have been obtained earlier.

$\binom{n-d-z}{r-1}$ is the number of ways of choosing a (positive) pool containing D_1 but no other D_j nor any unresolved negative. d times this quantity counts the number of ways of choosing a simple positive (not necessarily D_1), but no unresolved negatives. $\binom{n}{r} - \binom{n-d}{r}$ is the number of ways of choosing a positive row. Thus, the ratio

$$\frac{d\binom{n-d-z}{r-1}}{\binom{n}{r} - \binom{n-d}{r}}$$

gives the conditional probability that a positive pool contains a single positive and no unresolved negative (hence, the positive is identified). ■

Again, we derive P^+ independently.

Theorem 5.6:

$$P^+(D_1) = \sum_{u=0}^{n-d} \binom{n-d}{u} \sum_{j=u}^{n-d} (-1)^{j-u} \binom{n-d-u}{j-u}$$
$$\times \left[\frac{\binom{n}{r} - \binom{n-d}{r} - \binom{n-d-u}{r-1}}{\binom{n}{r}} + \frac{\binom{n-d-j}{r}}{\binom{n}{r}} \right]^t.$$

Proof: $\binom{n}{r} - \binom{n-d}{r}$ is the number of ways of choosing a positive pool. $\binom{n-d-u}{j-u}$ is the number of ways of choosing a pool containing D_1 but no other D_j nor an unresolved negative (D_1 is identified). Hence,

$$\binom{n}{r} - \binom{n-d}{r} - \binom{n-d-u}{r-1}$$

is the number of ways of choosing a positive pool not identifying D_1. $\binom{n-d-j}{r}$ is the number of ways of choosing a negative pool not containing any of the j specified negatives including $\bar{N}$. Theorem 5.6 now follows immediately from Theorem 2.19. ■

P^+ in Theorem 5.6 can be computed in $O(n^2)$ time.

Analytic solutions for optimal r to minimize either P^+ or $P(\bar{P} = 0)$ are not known.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

6. RANDOM DISTINCT *k*-SET DESIGN

RDkSD is neither column independent, nor row independent. Hence the computation of the probabilities of unresolved clones poses both a challenge but also an opportunity to expand the formulas beyond the independence threshold.

Lemma 6.1:

$$f(z) = \binom{t}{z} \sum_{i=z}^{t} (-1)^{i-z} \binom{t-z}{i-z} \frac{\binom{\binom{t-i}{k}}{d}}{\binom{\binom{t}{k}}{d}}.$$

Proof: All k appearances of a positive must be outside of Z. There are $\binom{t-i}{k}$ such distinct k-sets from which to choose d. Since the rows are not independent, the inclusion–exclusion formula is required. ■

Theorem 6.2:

$$P(\bar{N} = u) = \sum_{z=0}^{t} \binom{t}{z} \sum_{i=z}^{t} (-1)^{i-z} \binom{t-z}{i-z} \left[\frac{\binom{\binom{t-i}{k}}{d}}{\binom{\binom{t}{k}}{d}} \right]$$

$$\times \frac{\binom{\binom{t-z}{k}-d}{u}\binom{\binom{t}{k}-\binom{t-z}{k}}{n-d-u}}{\binom{\binom{t}{k}-d}{n-d}}.$$

Proof: There are $\binom{t-z}{k}$ k-sets not intersecting Z. d of them are chosen by the positives. The u unresolved negatives must be chosen from the remaining ones, and the $n-d-u$ resolved negatives must be chosen from the $\binom{t}{k} - \binom{t-z}{k}$ k-sets intersecting Z. Theorem 6.2 now follows from Theorem 2.4. ■

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

We argue for P^- independently.

THEOREM 6.3:

$$P^- = 1 - \sum_{i=1}^{k} (-1)^{i-1} \binom{k}{i} \frac{\binom{\binom{t-i}{k}}{d}}{\binom{\binom{t}{k}-1}{d}}.$$

PROOF: i is the number of rows intersecting C. The ratio represents the probability that no positive appears in these i rows. ■

We do not have a formula for $P(\bar{P} = v)$, even $f(z, y)$ seems too difficult to attempt. Hwang and Liu [5] gave formulas for P^+ and $f(z, y_1)$. Let $\epsilon_x = 1$ if $x = 0$, otherwise $\epsilon_x = 0$.

LEMMA 6.4:

$$f(z, y_1) = \binom{t}{z, y_1} \frac{\binom{t-z-y_1}{k-y_1} - \epsilon_{y_1}(d-1)}{\binom{t}{k} - (d-1)}$$

$$\times \sum_{h=0}^{t-z-y_1} (-1)^h \binom{t-z-y_1}{h} \frac{\binom{\binom{t-z-y_1-h}{k}}{d-1}}{\binom{\binom{t}{k}}{d-1}}.$$

PROOF: The sum in Lemma 6.4 gives this probability using the inclusion–exclusion formula, where $\binom{\binom{t-z-y_1-h}{k}}{d-1} \Big/ \binom{\binom{t}{k}}{d-1}$ is the probability that no $D_j, j \neq 1$, appears in a specified set of $z + y_1 + h$ pools (including the pools in $Z \cup Y_1$). There are $\binom{t-z-y_1}{k-y_1}$ ways of choosing D_1. However, if $y_1 = 0$, then D_1 is also chosen from the $\binom{t-z}{k}$ k-sets, hence $(d-1)$ k-sets, which have been selected as positives, should be subtracted. ■

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

Theorem 6.5:

$$P^+ = \sum_{z=0}^{t} \sum_{y_1=0}^{k} \binom{t}{z, y_1} \frac{\binom{t-z-y_1}{k-y_1} - \epsilon_{y_1}(d-1)}{\binom{t}{k} - (d-1)}$$

$$\times \sum_{h=0}^{t-z-y_1} (-1)^h \binom{t-z-y_1}{h} \frac{\binom{\binom{t-z-y_1-h}{k}}{d-1}}{\binom{\binom{t}{k}}{d-1}}$$

$$\times \sum_{i=0}^{y} (-1)^i \binom{y_1}{i} \frac{\binom{\binom{t}{k} - \binom{t-z}{k} + \binom{t-z-i}{k} - (d-1) - \epsilon_i}{n-d}}{\binom{\binom{t}{k} - d}{n-d}}.$$

Proof: $\binom{t}{k} - \binom{t-z}{k}$ is the number of k-sets intersecting Z, and $\binom{t-z-i}{k}$ is the number of k-sets intersecting neither Z nor the i specified rows. Thus, a k-set taken from the union of the two sets satisfies the condition in Theorem 2.16. However, the $(d-1)D_j$, $j \neq 1$, are also taken from the second set. Therefore, these $d-1$ k-sets must be subtracted before the $n-d$ negatives can be chosen. Further, if $i = 0$, then D_1 is also chosen from the second set; hence, one more k-set should be subtracted. ■

No analytic solution for optimal k to minimize any P^-, $P(\bar{N} = 0)$, P^+, and $P(\bar{P} = 0)$ is known.

7. SUMMARY AND NUMERICAL DATA

The method in [5] first computes the probability $u(j)$ that there are j unresolved negatives and then computes $f(z, y_1)$ from $\sum_j u(j) f(z, y_1 | j)$. The summation over j requires $O(n)$ times. The general approach we proposed in this article takes advantage of column independence in RID and RkSD to focus on the probability of a single negative blocking the identification of the positive and then to multiply that probability $(n-d)$-fold to account for all negative clones. Thus, there is no need to sum over j.

Bruno et al. [2] eliminated the summation over j for RkSD, not through the argument we offered but simply through combinatorial maneuvering. However, they overlooked something that resulted in an unnecessary inflation of the time complexity to $O(t^4)$. For RkSD and RDkSD, the range of y_1 is from 1 to k, where k is typically much smaller than t. Thus, the summation over y_1, as well as the summation over $O(y_1)$ terms when computing the probability that all y_1 appearances intersect with some unresolved negatives, should both involve $O(k)$ terms instead of $O(t)$. This brings a reduction of time complexity to $O(k^2t^2)$. Bruno et al. may have missed this point by using the variable $z + y_1$ instead of y_1, which somehow obscured the number of terms. We should also point out that the substitution of $O(k)$ for $O(t)$ in two summations in RkSD and RDkSD was also not observed in Hwang and Liu [5].

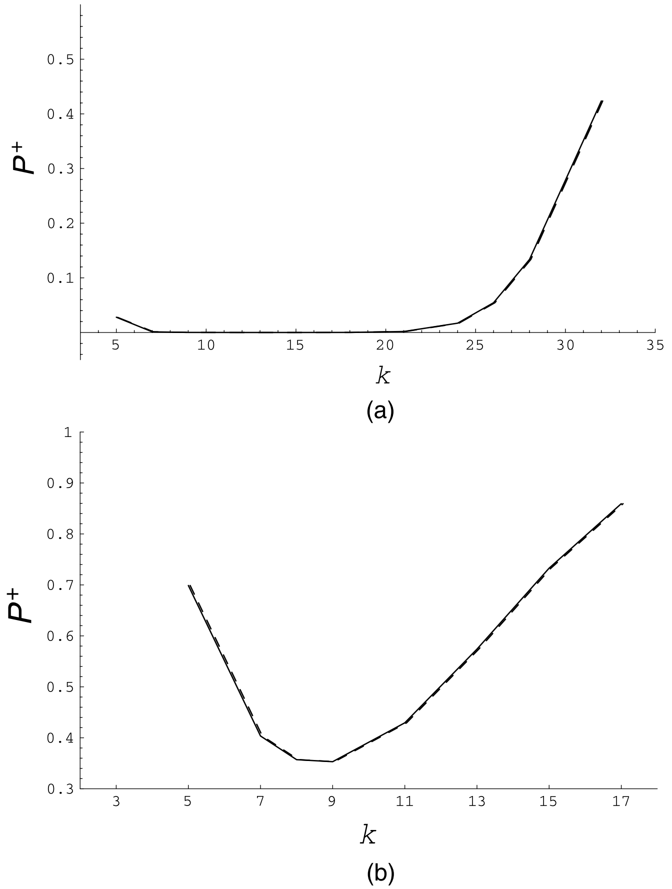

FIGURE 1. Comparison between RkSD and RDkSD (dashed line) with $n = 5000$, $t = 70$, and $d = 3$ (a) or $d = 5$ (b).

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

For RDkSD, although the columns are not independent, they are structured well enough so that we can argue over the $n - d$ negatives collectively—again, no need to introduce j. For RrSD, our general approach takes advantage of row independence to focus on the probability that a positive cannot be identified in a certain pool, and then to multiply that probability t-fold to account for all pools. Hence, the time complexity is reduced to $O(n^2)$, which is independent of t; the old method needs $O(n^2t^3)$ times.

The general approach helps us speed up the computation. We can only compute for $n \leq 100$ in [5], whereas we can compute for $n \geq 1000$, even for $n = 10{,}000$ for some designs now. Our program is written by Mathematica and not optimized. Hence, there is still the possibility for computation of larger parameters.

We present some numerical data in this section. First, we draw the RkSD and the RDkSD together in Figure 1 for easier comparison. As mentioned in Section 1, the

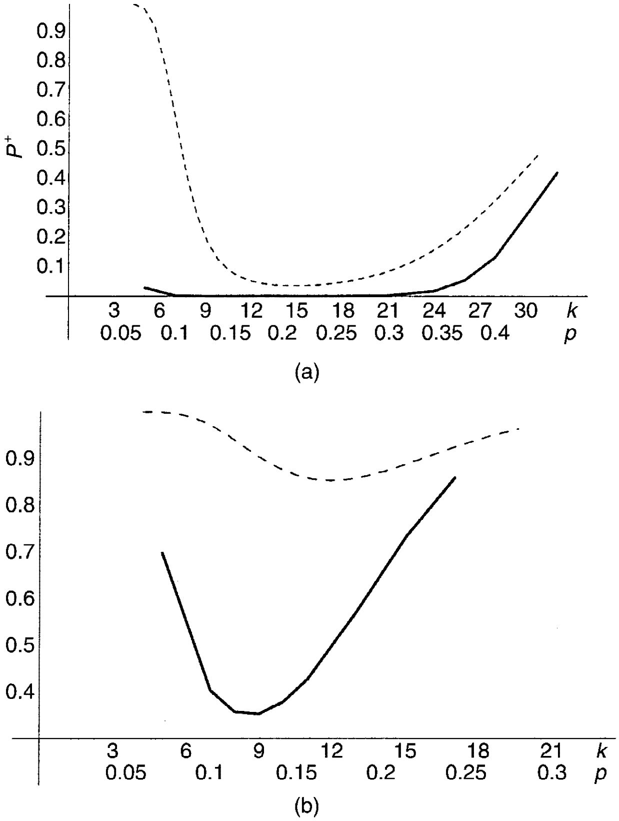

FIGURE 2. Comparison between RID (dashed line) and RkSD (solid line) with $n = 5000$, $t = 70$, and $d = 3$ (a) or $d = 5$ (b).

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

difference between RkSD and RDkSD is slight. Hence, during the pool's construction, rejecting any k-set that already occurs in the design [1] becomes unnecessary.

Then, the comparison of P^+ between RID and RkSD is presented in Figure 2. Because the range of possible p is from zero to one and the range of possible k is from zero to t, we make $k = t \times p$ for normalization. With the k restriction on the column weight, RkSD performs better than RID. Sometimes, the difference between these two designs is very significant and can be critical for their suitability. For example, in Figure 3a, when $n = 10{,}000$, $t = 85$, and $d = 5$, the optimal P^+ of RkSD is less than 0.1 and makes it a good design, whereas that of RID is about 0.69.

Figure 4 presents the comparison of P^+ between RID and RrSD. Here, we make $r = n \times p$. The performance of RrSD is about the same with RID, hence worse than RkSD. It seems to suggest that the column structure (of RkSD) is much more important than the row structure (of RrSD), although we have no explanation for it.

The data we show in Figure 4 has parameters $n = 1298$ and $t = 47$, which are much smaller than that we use in the comparison between RID and RkSD. This is because the time complexity of computing P^+ for RrSD is $O(n^2)$, whereas for RkSD, it is $O(k^2t^2)$. When n is large, kt is usually much smaller than n, so that P^+ of RkSD is still computable and it takes an unacceptably long time to compute P^+ for RrSD.

Although the time complexity of our formula for computing P^+ of RkSD is claimed to be $O(k^2t^2)$, which is independent of n, this ignores the fact that when n grows, the numbers in the formula have more bits and the division of large numbers takes longer to compute. Right now, we can compute for $n \leq 10{,}000$ with $kt \sim 1500$. We still need more efficient equations to deal with larger parameters (e.g., for n is in the order of 10^6).

In case that no explicit exact formulas can be obtained, we need good approximations in explicit forms. The reason of the need for explicit forms is not only for

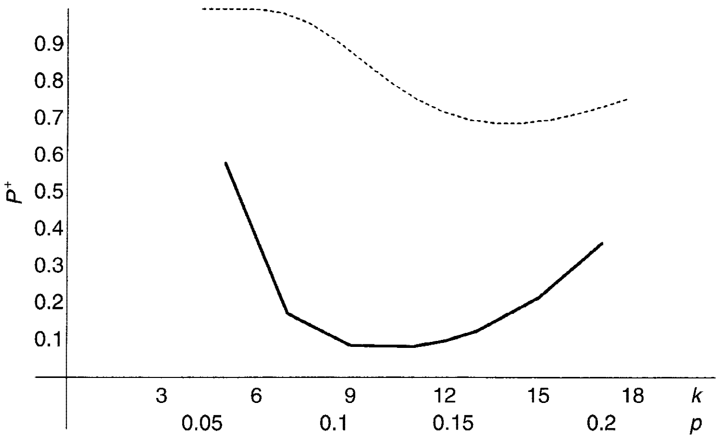

FIGURE 3. The performance difference between RID (dashed line) and RkSD (solid line) with $n = 10{,}000$, $t = 85$, and $d = 5$.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

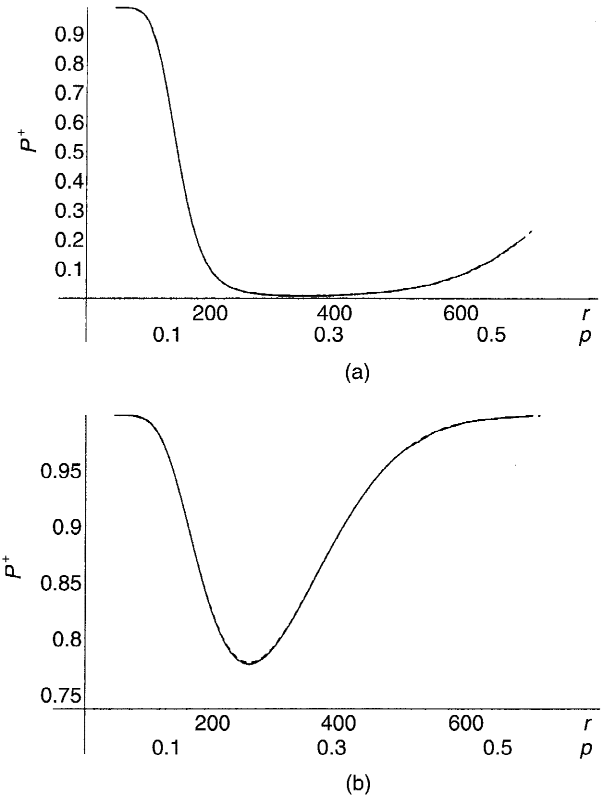

FIGURE 4. Comparison between RID (dashed line) and RkSD (solid line) with $n = 1298$, $t = 47$, and $d = 2$ (a) or $d = 4$ (b).

faster computation but also for being able to solve for optimal design parameters p, k, and r analytically. The numerical evidence certainly suggests that a unique optimum exists for each design.

Percus, Percus, Bruno, and Torney [9] gave an approximation whose leading term gives P^- if the rows were independent, then correction terms and some higher-order terms. For example, the approximation of P^- for RkSD is Eq. (42) of [9]. The first term is P^- for RkSD if the rows were independent. The second term corrects for this independence assumption and the third term reflects the consequences of dispersion and nondiscreteness of the number of positives.

References

1. Balding, D.J., Bruno, W.J., Knill, E., & Torney, D.C. (1995). *A comparative survey of non-adaptive pooling designs*, Genetic Mapping and DNA Sequencing, IMA Volumes in Mathematics and Its Applications. New York: Springer-Verlag, pp. 133–155.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press

2. Bruno, D.J., Knill, E., Balding, D.J., Bruce, D.C., Doggett, N.A., Sawhill, W.W., Stalling, R.L., Whittaker, C.C., & Torney, D.C. (1995). Efficient pooling designs for library screening. *Genomics* 26: 21–30.
3. Erdős, P.A., & Rényi, A. (1963). On two problems of information theory, *Magyar Tud. Akad. Mat. Kutató Int. Kőzl.* 8: 229–243.
4. Hwang, F.K. (2000). Random k-set pool designs with distinct columns. *Probability in the Engineering and Informational Sciences* 14: 49–56.
5. Hwang, F.K., & Liu, Y.C. (2001). The expected number of unresolved positive clones in various random pool designs. *Probability in the Engineering and Informantional Sciences* 15: 57–68.
6. Hwang, F.K., & Liu, Y.C. (to appear). Random pooling designs under various structures. *Journal of Combinatorial Optimization*.
7. Macula, A.J. (1997). A simple construction of *d*-disjunct matrices with certain weights. *Discrete Mathematics* 80: 311–312.
8. Macula, A.J. (1999). Probabilistic nonadaptive group testing in the presence of errors and DNA library screening. *Annals of Combinatorics* 1: 61–69.
9. Percus, J.K., Percus, O.E., Bruno, W.J., & Torney, D.C. (1999). Asymptotics of pooling designs performance. *Journal of Applied Probability* 36: 951–964.

https://doi.org/10.1017/S0269964804182028 Published online by Cambridge University Press