THE MAXIMUM EXPECTED NUMBER OF UNIQUE INDIVIDUALS IN A POPULATION

Ton de Waal*)

Abstract

A fundamental problem in the theory of statistical disclosure control is the determination of the probability that an individual with certain identifying features is unique in the population. One of the aims of statistical disclosure is to avoid the publication of records of these unique individuals. In this paper an upper bound for the probability of uniqueness is derived. For this purpose the problem is stated in terms of urns and balls. The resulting optimization problem is solved by means of elementary and numerical analysis.

Keywords: statistical disclosure, optimization, urn models

---

*) Statistics Netherlands, Department of Statistical Methods, P.O. Box 959, 2270 AZ Voorburg, The Netherlands, (070-3374930).

The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

1. Introduction

A central problem in the theory of statistical disclosure is the determination of the probability that an individual with certain identifying features is unique in the population. The reason for this is that individuals with unique identifying features are (relatively) easy to recognize. Therefore, to estimate the risk of disclosure that is involved when microdata are released it is useful to estimate the number of unique individuals in the population. In this paper an upper bound for the number of unique individuals in the population is derived for a given disclosure key.

For convenience the problem is stated in terms of urns and balls instead of identifying features and individuals. Individuals correspond to balls and identifying features to urns. Each individual of the population has a probability p_j to have identifying feature j (i.e. a specific score on a disclosure key). In terms of balls and urns: each ball has a probability p_j to be assigned to urn j. The balls are assigned to the urns independently. The problem is to find the probability distribution for which the expected number of individuals with unique identifying features is maximal. In other words we want to find the probability distribution for which the expected number of urns with exactly one ball is maximal.

In Section 2 of this paper the problem is stated. In Section 3 some consequences of the Lagrangian are examined. Section 4 and Section 5 are rather technical. From these two sections an important result follows, namely that there are (at most) two possible solutions. In Section 6 bounds are derived in order to be able to compare the results from the two possible solutions without explicitly determining them. Numerical results are presented in Section 7. A short summary of the solution obtained is given in Section 8.

2. The problem

Suppose we have m urns and n balls. Each ball is assigned to an urn independently. The probability to assign a ball to the j-th urn is p_j. The expected number of urns with exactly one ball can be expressed as a function of the probabilities p_j. We are interested in the following problem: "How should the probabilities p_j be chosen in order to maximize the expected number of urns with exactly one ball ?"

In Section 1 we already noted that this problem is equivalent to a problem in the theory of statistical disclosure. If we let the possible identifying features correspond to the urns and the individuals to the balls, then we see that solution of our problem provides an upper bound for the number of unique individuals in the population. Here we assume that the

identifying features are distributed independently. This assumption is not always justified in practice.

It is easy to calculate the expected number of urns with exactly one ball. This number is given by

$$E = \sum_{j=1}^{m} np_j(1 - p_j)^{n-1}. \tag{1}$$

In the rest of this paper we will also use the function N defined by

$$N = E/n$$

We will refer to each of these funtions as the target function. We hope that this will not confuse the reader too much. The p_j's must be larger than, or equal to, zero. They must also sum to unity. These constraints are expressed by

$$p_j \geq 0 \quad , \text{ for all } j = 1, 2, \ldots, m, \tag{2}$$

$$\sum_{j=1}^{m} p_j = 1. \tag{3}$$

3. Consequences of the Lagrangian

In order to find the maximum of (1) subject to (2) and (3) we begin by determining the Lagrangian $L(p_1,\ldots,p_m,\lambda)$.

$$L(p_1,\ldots,p_m,\lambda) = \sum_{j=1}^{m} np_j(1 - p_j)^{n-1} - \lambda(\sum_{j=1}^{m} p_j - 1) \tag{4}$$

By differentiating L with respect to λ we obtain that the sum of the p_i's is equal to one. By differentiating with respect to p_i we obtain

$$n(1 - p_i)^{n-1} - n(n - 1)p_i(1 - p_i)^{n-2} = \lambda. \tag{5}$$

This must hold for all i. Therefore we can conclude that the optimal p_i's obey

$$(1 - np_i)(1 - p_i)^{n-2} = (1 - np_j)(1 - p_j)^{n-2} \tag{6}$$

This relation must hold for all i and j.

Relation (6) suggests that it is useful to study the behaviour of the function defined by

$$f_n(x) = (1 - nx)(1 - x)^{n-2} \quad , 0 \leq x \leq 1. \tag{7}$$

We can make the following observations about this function :

1. $f_n(0) = 1$
2. $f_n(1) = 0$
3. $f_n(1/n) = 0$
4. $f'_n(x) = (n - 1)(nx - 2)(1 - x)^{n-3}$
5. $f'_n(x) = 0$ if $x = 2/n$ or $x = 1$
6. $f'_n(x) < 0$ if $x < 2/n$
7. $f'_n(x) > 0$ if $2/n < x < 1$

This implies that the equation $f_n(x) = C$ has the following solutions for $0 \leq x \leq 1$:

8. If $0 < C \leq 1$, then there is only one solution. For this solution x_0 we have: $0 \leq x_0 < 1/n$. In order words, for $0 < x < 1/n$ the function f is injective.

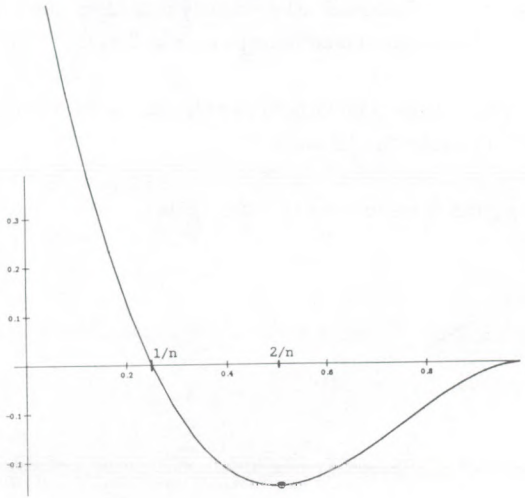
9. If $f_n(2/n) < C \leq 0$, then there are two solutions x_1 and x_2 between 0 and 1. For these solutions we have: $1/n \leq x_1 < 2/n$ en $2/n < x_2 \leq 1$. In other words, for $1/n < x < 1$ the function f is not injective.

10. If $C = f_n(2/n)$, then there is only one solution: $x = 2/n$.

This reveals that the optimal p_i's can have at most two different values. Moreover, we know that when the optimal solution has two different p_i-values, then one value lies between 1/n and 2/n and the other is larger than 2/n. We also know that if one p_i is smaller than 1/n, then all the p_i's have the same value. This implies that if $m \geq n$ then the optimal solution is given by $p_i = 1/m$ for all i. From now on we therefore assume that $n > m$.

In order to have a visual "proof" of the observations 8, 9 and 10 we have plotted function f. In Figure 1 the function f is drawn for the case $n = 4$. From Figure 1 one can clearly see that observations 8, 9 and 10 hold in this case.

Figure 1. The plot of the graph of the function $f_4(x) = (1 - 4x)(1 - x)^2$



4. Parametrisation of conjugated values

We know that the solutions of the set of equations given by (6) have at most two different values for $0 \leq p_i \leq 1$. From now on we will call these two values conjugated values. In this section a parametrisation of conjugated values is derived. To simplify the notation somewhat we use $1 - p_i$ instead of p_i in this section.

Suppose we have two conjugated values $p_1 = 1 - Q$ and $p_2 = 1 - R$. We suppose that $R = \mu Q$. From the set of equations (6) we obtain relation (8) between R and Q:

$$nQ^{n-1} + (1-n)Q^{n-2} = nR^{n-1} + (1-n)R^{n-2} \qquad (8)$$

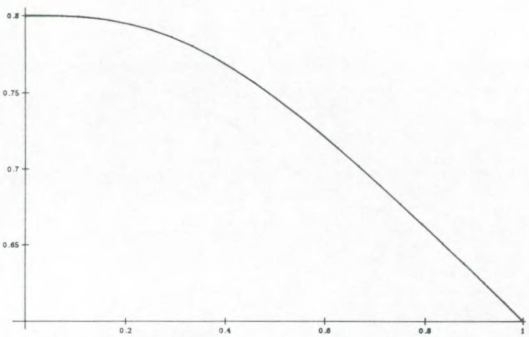
If we substitute $R = \mu Q$ in (8), then we find a parametrisation of Q in terms of μ.

$$Q = \frac{(n - 1)}{n} \frac{(1 - \mu^{n-2})}{(1 - \mu^{n-1})}, \quad 0 \leq \mu < 1 \qquad (9)$$

If $\mu = 0$, then Q is equal to (n-1)/n. The associated probability p_1 is therefore equal to 1/n. R is equal to 0 if $\mu = 0$. The associated probability p_2 is equal to 1. When μ approaches 1 then Q tends to (n-2)/n. The associated probability p_1 tends to 2/n. R tends to (n-2)/n when μ approaches 1. The associated probability p_2 tends to 2/n.

The behaviour of function Q is maybe a bit difficult to understand without some visual aid. In Figure 2 the function Q is drawn for the case n = 5.

Figure 2. The plot of the graph of the function $Q = 4(1 - \mu^3)/5(1 - \mu^4)$



The derivative of Q with respect to μ is given by

$$\frac{dQ}{d\mu} = \frac{(n - 1)}{n} \frac{(n - 1)(1 - \mu^{n-2})\mu^{n-2} - (n - 2)(1 - \mu^{n-1})\mu^{n-3}}{(1 - \mu^{n-1})^2}. \qquad (10)$$

This derivative is less than 0 if $0 \leq \mu \leq 1$. The parametrisation of Q and R by means of μ is therefore 1-1. By the way, to see that the derivative of Q with respect to μ is less than zero we do not have to calculate this derivative explicitly, but we only have to look at the function f. The derivative of R with respect to μ is larger than 0 if $0 \leq \mu \leq 1$.

5. The number of urns with the same probability

5.1 An important set of equations involving conjugated values

Now we make use of the fact that there are at most two (conjugated) values p and q for the optimal probabilities. We suppose that there are z urns with probability p and m - z urns with probability q. Implicitely we hereby assume that $n > m$ and therefore (cf. points 8 and 9 of Section 3) that both probabilities are larger than 1/n. Note that $z = m$, or $z = 0$, corresponds with the situation that all probabilities are equal. As the reader will remember we have already established that the optimal probabilities are all equal to 1/m in case $n \leq m$.

Now we will try to find the optimal value of z for the target function N. We do this by substituting z and the optimal conjugated values $p = p(z)$ and $q = q(z)$ in the target function N. This gives us another function which will be denoted by M(z). For p and q relation (6') and relation (11) hold.

$$(1 - p)^{n-2}(1 - np) = (1 - q)^{n-2}(1 - nq) \qquad (6')$$

and

$$zp + (m - z)q = 1. \qquad (11)$$

For the moment we do not demand z to be an *integer* between 0 and m. Instead z may be any *real* value between 0 and m. We will first solve the problem for z assumed to be a real value, and later we will modify this solution to obtain the solution for z being integer.

5.2 The solutions of the equations

In this section we show that the set of equations (6') and (11) has at most three different pairs of solutions $(p_i(z),q_i(z))$. These pairs are differentiable with respect to z. The proof of this statement is elementary, but rather long and tedious. Therefore, we do not go into all the details of the proof.

The first solution is, of course, given by $p(z) = q(z) = 1/m$. From now on we concentrate on the case that p(z) is unequal to q(z). Without loss of generality we assume that $p(z) < q(z)$. Instead of relation (11) we use the following relation

$$z(1 - p) + (m - z)(1 - q) = m - 1. \tag{12}$$

For (1 - p) we can substitute the expression given in (9), and for (1 - q) we can substitute μ times that expression. So, the problem of finding solutions to the set of equations (6') and (11) translates into the problem of finding the roots of the function $h(\mu)$ defined by

$$h(\mu) = z\,\frac{n-1}{n}\,(1 - \mu^{n-2}) + (1-z)\,\frac{n-1}{n}\,\mu(1 - \mu^{n-2}) - (m-1)(1 - \mu^{n-1}) \tag{13}$$

for $0 \leq \mu < 1$. Now we will list some properties of the function h.

It is easy to see that for $\mu = 0$ and $\mu = 1$ we have

1. $h(0) = z(n - 1)/n - (m - 1)$
2. $h(1) = 0$

The function $h(\mu)$ can be studied by examining its first and second derivative. The first derivative is still a complicated expression, but for $\mu = 0$ and $\mu = 1$ we obtain two simple terms:

3. $h'(0) = (m - z)(n - 1)/n$
4. $h'(1) = (2m - n)(n - 1)/n$

The second derivative is given by:

5. $h''(\mu) = (n-2)(n-1)((m-1)/m-(1-z)(n-1)/n)\mu^{n-3} - ((n-3)(n-2)(n-1)z/n)\mu^{n-4}$

So, it is very easy to determine when $h''(\mu)$ is positive and when it is negative.

Combining these, and other, facts about the function $h(\mu)$ we are able to draw the following conclusions.

If $n \geq 2m$, then $h(\mu)$ has one root between 0 and 1
If $n < 2m$, then $h(\mu)$ has at most two roots between 0 and 1

The first case is not very interesting. We can make the remark, however, that the pair $(p(z),q(z))$ associated to the root $\mu(z)$ is the optimal solution for given value of z. To see

this we can consider the target function N(p,q), which is of course defined by

$$N(p,q) = (m - 1)p(1 - p)^{n-1} + q(1 - q)^{n-1}, \qquad (14)$$

where $q = 1 - (m-1)p$. We are seeking the maximum of this function. One possible solution is the pair (p(z),q(z)) associated to $\mu(z)$. The second derivative of N(p,q) with respect to p is given by

$$\frac{\partial^2 N}{\partial p^2} = (m-1)(n-1)((np - 2)(1 - p)^{n-3} + (m-1)(nq - 2)(1 - q)^{n-3}) \qquad (15)$$

This function is positive for the only other possible solution, $p = q = 1/m$, in case $n > 2m$. Therefore, the target function has a local minimum for $p = q = 1/m$. So, the only remaining possible solution for the maximum is the pair (p(z),q(z)) associated to $\mu(z)$.

We will describe the second case, $n \leq 2m$, in more detail. If there is at least one root, then there is one root $\mu(z)$ for which the associated p(z) converges to 1/m when z tends to m. If there is a value z_0 for which there are two roots, then there is one root $\mu_1(z)$ which exists for all $z_0 \leq z \leq m$, while the other root $\mu_2(z)$ exists for $z_0 \leq z \leq n(m-1)/(n-1)$, but not for $z > n(m-1)/(n-1)$. This root μ_2 is equal to 0 for z equal to n(m-1)/(n-1). For larger values of z μ_2 would become smaller than 0, which is not allowed.

By applying the "implicit function theorem" we can show that for each value of z between 0 and m for which they exist the pairs (p(z),q(z)) are differentiable with respect to z. We only have to determine the determinant of the Jacobian of the set of equations (6') and (11). This determinant has to be unequal to zero in order to be able to apply the implicit function theorem. After some, not too difficult, calculations it becomes clear that the determinant is indeed unequal to zero.

5.3 Implications of the solutions of the important set of equations

Now we will use the (differentiable) functions p(z) and q(z) to determine the possible optimal values of z. We substitute the pair of functions p(z) and q(z) into the target function N. This gives us another function M(z). The function M is given by

$$M(z) = zp(1 - p)^{n-1} + (m - z)q(1 - q)^{n-1}. \qquad (16)$$

Because relation (11) is valid for all z, we can differentiate with respect to z. We arrive at the following result.

$$(m - z)\frac{dq}{dz} = q - p - z\frac{dp}{dz} \tag{17}$$

Now we are able to determine the derivative of M with respect to z. This will enable us to deduce the possible optimal values for z for the target function N. Finally, we have to compare the possible optimal values of N to find the true optimal value. By applying the chain rule, relation (6') and relation (17) we can show that the derivative of M with respect to z is given by (18).

$$\frac{dM}{dz} = \left[(n-1)\,(p^2(1 - p)^{n-2} - q^2(1 - q)^{n-2})\right]\Big|_z \tag{18}$$

Using (6'), or equivalently relation (6), again we find relation (19).

$$(1 - q)^{n-2} = (1 - p)^{n-2}\frac{1 - np}{1 - nq} \tag{19}$$

When we substitute this into (18) we finally arrive at

$$\frac{dM}{dz} = \left[\frac{(n - 1)(1 - p)^{n-2}}{1 - nq}(p - q)(p + q - npq)\right]\Big|_z \quad . \tag{20}$$

Now we have succeeded in finding an expression for the derivative of M with respect to z. We can therefore determine the optimal z. This turns out to be very easy, because dM/dz has a fixed sign.

Without loss of generality we assume that $p < q$. This is of course equivalent to: $1 - p > 1 - q$. We know that $p - q < 0$ and $1 - nq < 0$ (see the conclusions of Section 3). So, to establish the sign of dM/dz we only have to determine the sign of $p + q - npq$. This may seem a hard problem, because p and q both depend on the value of z. However, by using the parametrisation of p and q we can demonstrate that the sign of dM/dz is independent of the actual value of z.

We can rewrite $p + q - npq$ to obtain

$$p + q - npq = -n(1-p)(1-q) + (n-1)(1-p) + (n-1)(1-q) + (2-n). \tag{21}$$

From (21) and the parametrisation of $1 - p$ and $1 - q$ we can derive

$$p + q - npq = -\frac{1}{(1 - \mu^{n-1})^2}\frac{n-1}{n}(F(\mu) - G(\mu)). \tag{22}$$

Here $F(\mu)$ and $G(\mu)$ are given by

$$F(\mu) = (1 - \mu^{n-1})((n - 1)\mu(1 - \mu^{n-2}) - \frac{n(n - 2)}{(n - 1)}(1 - \mu^{n-1})), \tag{23}$$

$$G(\mu) = (1 - \mu^{n-2})((n - 1)\mu(1 - \mu^{n-2}) - (n - 1)(1 - \mu^{n-1})). \tag{24}$$

Because $1 - \mu^{n-1} > 1 - \mu^{n-2}$ and $n(n - 2)/(n - 1) < n - 1$, we find that $F(\mu)$ is larger than $G(\mu)$ when $0 \leq \mu < 1$. In other words, $p + q - npq$ is larger than 0 when $0 \leq \mu \leq 1$. This result is independent from z. Therefore we can conclude that dM/dz is larger than 0. This implies that in order to optimize the function M we have to make z as large as possible. This, in turn, implies that in order to optimize the target function N we have to make z as large as possible. The actual optimal (real) value of z is determined by the constraints, but we know that the largest value of z that satisfies all the constraints is the optimal value.

So far we have allowed z to be any real number between 0 and m. Now we remind ourselves that z must be an integer between 0 and m. We know that if $p < q$, then z must be as large as possible.

We have the following cases:

a. If $n \leq m$, then the optimal solution is the uniform distribution.

b. If $n > 2m$, then the uniform distribution is a local minimum. The optimal real value for z is $n(m-1)/(n-1)$, which is larger than $(m-1)$, but smaller than m. Therefore, the optimal solution is given by a non-uniform distribution with $(m-1)$ small probabilities and one large probability.

c. If $m < n \leq 2m$, then there are two possibilities: either the optimal solution is the uniform distribution or the optimal solution is a non-uniform distribution with $(m-1)$ small probabilities and one large probability. However, it is not clear for which combinations of m and n the uniform distribution is the optimal solution and for which combinations of m and n the non-uniform distribution is the optimal solution. In Section 6 this case, i.e. $m < n \leq 2m$, is further investigated.

6. General remarks about the solution

The solution for $n \leq m$ is given by p_i is equal to $1/m$ for all i. For $n > 2m$ the solution has m-1 probabilities smaller than $2/n$ and one probability larger than $2/n$. In case $m < n \leq 2m$ the solution is not clear. In this section we make some remarks about this case.

We start by making the observation that if the non-uniform distribution, i.e. (m-1) probabilities equal to p ($1/n < p < 2/n$) and one probability equal to q ($2/n < q < 1$; $q = 1 - (m-1)p$), is better than the uniform distribution for a certain number of balls n_0, then this non-uniform distribution is better than the uniform distribution for all $n \geq n_0$. The proof of this assertion is quite simple. To make the dependency on n more explicit we use the notation:

$$N(p,q;n_0) = (m - 1)p(1 - p)^{n_0-1} + q(1 - q)^{n_0-1} \tag{25}$$

Now, let us suppose that for a certain n_0 the non-uniform distribution is better than the uniform distribution. In other words, we have the following relation

$$N(p,q;n_0) \geq (1 - 1/m)^{n_0-1} \tag{26}$$

We have to prove that a similar inequality for $n_0 + 1$ instead of n_0 holds. We can do this by making use of the inequality for n_0, and by rewriting the expression for $N(p,q;n_0+1)$. So, we write $N(p,q;n_0+1)$ in the following way

$$\begin{aligned} N(p,q;n_0+1) = {} & N(p,q;n_0)(1-1/m) + (m-1)p(1/m - p)(1 - p)^{n_0-1} \\ & + q(1/m - q)(1 - q)^{n_0-1} \end{aligned} \tag{27}$$

We can combine the last two terms of this expression by making use of another inequality, namely

$$p(1 - p)^{n_0-1} \geq q(1 - q)^{n_0-1} \tag{28}$$

This inequality holds because $1/(n_0+1) < 1/n_0 < p < q$. Using this inequality to combine the last two terms of the expression for $N(p,q,n_0+1)$ we find yet another inequality, namely

$$N(p,q;n_0+1) \geq (1 - 1/m)^{n_0} + \left[1 - (m-1)p - q \right] q(1 - q)^{n_0} \tag{29}$$

Finally, by using 1 - (m-1)p - q = 0, we see that we have succeeded in deriving the desired inequality. So, we can draw the conclusion that if, for a certain n_0, there is a non-uniform distribution that is better than the uniform distribution, then for all $n \geq n_0$ there is a non-uniform distribution that is better than the uniform distribution. The problem remains to determine the critical number n_0, given a certain number of urns m.

We can evaluate the target function N by assuming that the solution is given by uniformly distributed p_i, i.e. $p_i = 1/m$. The value of N for uniformly distributed p_i is denoted by N_{uni}. For N_{uni} we have the following expression.

$$N_{uni} = \left(\frac{m - 1}{m}\right)^{n-1} \tag{30}$$

By assuming that the non-uniform solution of the equation for the conjugated values exists we can estimate the value of the target function for this solution. This value will be denoted by N_{non}, the maximal expected number of urns with exactly one ball in case of a non-uniform distribution. In Section 5 we have derived that the target function is maximal if z is as large as possible. The largest possible real value for z is n(m-1)/(n-1). Therefore, an upper bound for N_{non} is given by

$$N_{non,max} = \frac{m - 1}{n - 1}\left(\frac{n - 1}{n}\right)^{n-1} \tag{31}$$

On the other hand we can find a lower bound $N_{non,low}$ for N_{non}. This can be obtained by substituting any probability distribution p_i into the target function N. Very suitable is:

$p_i = 1/n$, for i = 1, 2, ..., m - 1
$p_m = 1 - (m-1)/n$

Substituting this in N yields

$$N_{non,low} = \frac{m - 1}{n}\left(\frac{n - 1}{n}\right)^{n-1} + \left(1 - \frac{m - 1}{n}\right)\left(\frac{m - 1}{n}\right)^{n-1} \tag{32}$$

Relations (30), (31) and (32) give criteria to decide which distribution is better:

if $N_{uni} \geq N_{non,max}$, then the uniform distribution is better
if $N_{uni} \leq N_{non,low}$, then the non-uniform distribution is better.

In these two cases we can decide which distribution will be the optimal one without

explicitly determining the non-uniform distribution. We are able to decide which distribution is the optimal distribution immediately by looking at the numbers m and n. By the way, if the latter situation occurs, i.e. if $N_{uni} \leq N_{non,low}$, then we can be sure that the non-uniform distribution exists. If N_{uni} lies between $N_{non,low}$ and $N_{non,max}$, then we have to determine the non-uniform distribution and substitute this solution into the target function. In this case we cannot decide which distribution will be best without explicitly determining the non-uniform distribution.

We can investigate the behaviour of N_{uni} and N_{non} if n tends to infinity. We can do this for two different cases. Firstly, we can assume that $m = \alpha n$, where $0 < \alpha < 1$ is a constant. Secondly, we can assume that $m = n - ß$, where ß is a constant.

If $m = \alpha n$, then we have

$$\lim_{n\to\infty} N_{uni} = e^{-1/\alpha} \tag{33}$$

and

$$\lim_{n\to\infty} N_{non,max} = \lim_{n\to\infty} N_{non,low} = \alpha/e \tag{34}$$

It is an elementary exercise to check that

$$e^{-1/\alpha} < \alpha/e \tag{35}$$

for $0 < \alpha < 1$. So, we can conclude that if we keep the ratio $\alpha = n/m$ fixed, then the non-uniform distribution is always better than the uniform distribution for n large enough.

If $m = n - ß$, then we have

$$\lim_{n\to\infty} N_{uni} = \lim_{n\to\infty} N_{non,max} = \lim_{n\to\infty} N_{non,low} = 1/e \tag{36}$$

In fact we have the following

$$(n - 1)(N_{uni} - N_{non,max}) = \frac{ß}{e} + g(n) \text{ , where } \lim_{n\to\infty} g(n) = 0 \tag{37}$$

So, we can conclude that if we keep the difference $ß = n - m$ fixed, then the uniform distribution is better than any non-uniform distribution for n large enough.

7. Numerical results

Given the number of urns m it is interesting to know the smallest number of balls for which the solution is not uniformly distributed. As a first step to find this number we can apply the criteria given in Section 6 to determine which distribution is better. In Table 1 the highest number of balls for which E_{uni} is still larger than $E_{non,max}$, n_u, is listed. For this number of balls and for any smaller number of balls the uniform distribution is the optimal one. In Table 1 the lowest number of balls for which E_{uni} is smaller than $E_{non,low}$, n_n, is also listed. For this number of balls and for any larger number of balls the non-uniform distribution is the optimal one. For numbers of balls between n_u and n_n the criteria given in Section 6 cannot determine which distribution is the optimal distribution.

Table 1. Critical values n_u and n_n as given by the criteria from Section 6.

number of urns m	n_u	n_n
5	6	9
10	11	16
15	16	22
20	21	27
30	31	39
40	41	50
50	51	61
60	61	72
70	71	83
80	81	94
90	91	105
100	101	115
150	151	168
200	201	221
300	301	326
400	401	429
500	501	533
1 000	1 001	1 046

For numbers of balls between n_u and n_n it is not clear yet which distribution is the optimal one. Therefore a numerical routine has been implemented. This routine determines the solution of the problem given m urns and n balls, and computes the expected number E of urns with exactly one ball. The results of some numerical experiments are shown in Table 2. In this table the "last optimal uniform distribution" and the "first non-uniform optimal distribution" are listed. By this we mean that for any number of balls smaller than, or equal to, the number of balls of the last uniform optimal distribution the uniform distribution is the optimal distribution. For any number of balls larger than, or equal to,

the number of balls of the first non-uniform distribution the non-uniform distribution is the optimal distribution. In other words the number of balls of the first non-uniform distribution is the critical value $n_0(m)$. In case of the uniform distribution the probabilities are, of course, given by 1/m. In case of the non-uniform distribution there are m-1 small probabilities p and one large probability given by 1 - (m-1)p. The small probability p is also listed in Table 2.

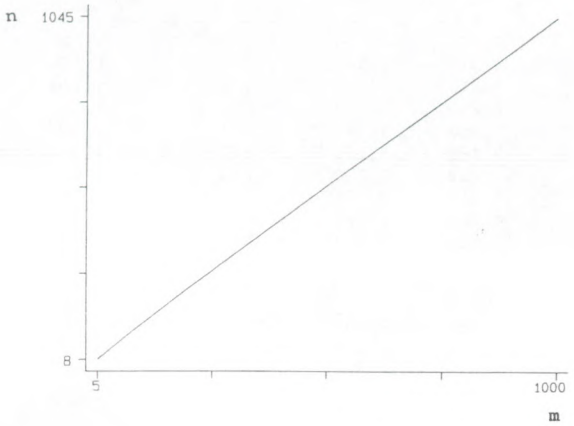
Table 2. The solution of the problem for given m and n

number of urns (m)	last optimal uniform distribution		first optimal non-uniform distribution		
	#balls (n)	value target function E	#balls (n)	value target function E	smallest p
5	8	1.678	9	1.568	1.16x 0.1
10	15	3.432	16	3.420	6.28x 0.01
15	21	5.284	22	5.271	4.56x 0.01
20	26	7.212	27	7.123	3.71x 0.01
30	38	10.840	39	10.807	2.57x 0.01
40	49	14.535	50	14.493	2.00x 0.01
50	60	18.218	61	18.175	1.64x 0.01
60	71	21.893	72	21.816	1.39x 0.01
70	82	25.565	83	25.538	1.20x 0.01
80	93	29.235	94	29.218	1.06x 0.01
90	104	32.902	105	32.898	9.52x 0.001
100	114	36.617	115	36.579	8.70x 0.001
150	167	55.016	168	54.978	5.95x 0.001
200	220	73.397	221	73.374	4.52x 0.001
300	325	110.170	326	110.165	3.07x 0.001
400	428	146.979	429	146.955	2.55x 0.001
500	532	183.751	533	183.744	1.88x 0.001
1 000	1 045	367.694	1 046	367.687	9.56x0.0001

From Table 1 and Table 2 we see that an efficient practical way to determine which distribution is the optimal distribution is comparing the numbers E_{uni} and $E_{non,low}$. If $E_{non,low}$ is larger than E_{uni}, then the non-uniform distribution is the optimal one. In this case the optimal value of E is almost equal to $E_{non,low}$. If E_{uni} is larger than $E_{non,low}$, then the uniform distribution is the optimal one. In this case the optimal value of E is equal to E_{uni}.

In Figure 3 it is shown for which combinations of m and n the uniform distribution is the optimal one, and for which combinations of m and n a non-uniform distribution is the optimal one.

Figure 3. The number of balls of the first optimal non-uniform distribution as a function of the number of urns.



The graph of the number of balls of the first optimal non-uniform distribution as a function of the number of urns is almost a straight line. This is, of course, not very surprising if we consider Table 2. However, the result is rather surprising if we look at the complexity of the equations which describe the relation between the number of urns, the number of balls and the optimal distribution.

In Table 3 the expected number of urns with exactly one ball for the two possible optimal distributions are compared. In this table some as yet undefined quantities, E_{uni}, E_{non}, $E_{non,low}$ and $E_{non,max}$, are used. They are defined, of course, by n times the corresponding N-value. In case there are 5 urns and 8 balls the possible optimal non-uniform distribution does not exist, because the function $h(\mu)$ does not have a root between 0 and 1. Notice that the value of E_{non} is extremely well approximated by $E_{non,low}$. So, in practice we may use $E_{non,low}$ instead of E_{non}.

Table 3. Comparison between the two possible optimal distributions

# urns	# balls	E_{uni}	E_{non}	$E_{non,low}$	$E_{non,max}$
5	8	1.678	-	1.602	1.795
5	9	1.510	1.568	1.567	1.754
10	15	3.432	3.431	3.430	3.670
10	16	3.294	3.420	3.420	3.646
15	21	5.284	5.279	5.279	5.540
15	22	5.166	5.271	5.271	5.522
20	26	7.212	7.130	7.130	7.412
20	27	7.115	7.123	7.123	7.396
30	38	10.840	10.811	10.811	11.103
30	39	10.754	10.807	10.807	11.092
40	49	14.535	14.496	14.496	14.797
40	50	14.461	14.493	14.493	14.788
50	60	18.218	18.178	18.178	18.486
50	61	18.151	18.175	18.175	18.478
60	71	21.893	21.859	21.859	22.171
60	72	21.832	21.857	21.857	22.165
70	82	25.565	25.540	25.540	25.855
70	83	25.507	25.538	25.538	25.849
80	93	29.235	29.220	29.220	29.537
80	94	29.180	29.218	29.218	29.532
90	104	32.902	32.900	32.900	33.219
90	105	32.850	32.898	32.898	33.214
100	114	36.617	36.581	36.581	36.904
100	115	36.850	36.579	36.579	36.900
150	167	55.016	54.979	54.979	55.310
150	168	54.976	54.978	54.978	55.307
200	220	73.397	73.375	73.375	73.710
200	221	73.362	73.374	73.374	73.708
300	325	110.170	110.165	110.165	110.505
300	326	110.140	110.165	110.165	110.504
400	428	146.979	146.956	146.956	147.300
400	429	146.954	146.955	146.955	147.299
500	532	183.751	183.745	183.745	184.091
500	533	183.728	183.744	183.744	184.090
1 000	1 045	367.693	367.688	367.688	368.040
1 000	1 046	367.677	367.687	367.687	368.039

8. Summary

The first result we derived was that the probabilities of the solution can have at most two different values. This was a rather easy result, obtained in Section 3.

We still did not know how many probabilities have one of the possible values and how many probabilities have the other possible value, though. This question was examined in Section 4 and Section 5. After much ado, we found that there are two possibilities: either all the probabilities are equal, or there are m-1 small probabilities, which are all equal, and one large probability.

At that moment we were faced with the question of determining for what combinations of m and n all the optimal probabilities have the same value, and for what combinations of m and n there are m-1 small probabilities and one large probability. Part of the answer to this question was already obtained in Section 5. In case $n \leq m$ the optimal probabilities are all equal. In case $n \geq 2m$ there is one large probability and m-1 small probabilities. In Section 6 we examined this question in more detail. We were able to describe the behaviour of the optimal solution in case m and n tend to infinity if we assume that either the ratio, or the difference, of m and n is fixed. We also obtained an upper bound, and a lower bound, for the expected number of urns with exactly one ball for the possibly optimal non-uniform distribution. This gives us a criterion to decide whether the uniform distribution or the non-uniform distribution is better, without explicitly determining the possibly optimal non-uniform distribution.

If $N_{uni} \geq N_{non,max}$, then the uniform distribution is better,
if $N_{uni} \leq N_{non,low}$, then the non-uniform distribution is better.

Unfortunately, there are still combinations of m and n for which we are unable to determine which distribution is the optimal distribution without explicitly determining the possibly optimal non-uniform distribution. For these cases we have $N_{non,low} \leq N_{uni} \leq N_{non,max}$. In Section 7 numerical results were presented for a number of cases. From these numerical results we can conclude that in order to determine the optimal distribution it is in general sufficient to compare the numbers N_{uni} and $N_{non,low}$. If $N_{uni} > N_{non,low}$, then the uniform distribution is the optimal one. If $N_{non,low} > N_{uni}$, then the non-uniform distribution is the optimal one.

Ontvangen: 7-9-1993
Geaccepteerd: 11-2-1994