Scientific Research Publishing

Applied Mathematics, 2020, 11, 1253-1274
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

A Stochastic Strategy for Optimizing the Cost of Systematic Screening of Some Diseases in a Situation of Insufficient Equipment

Lucien Zihindula Biguru[1,2]

[1]Faculty of Sciences, Université Catholique de Bukavu (UCB), Bukavu, Congo
[2]Faculty of Sciences, Université Officielle de Bukavu (UOB), Bukavu, Congo
Email: lucien.zihindula@ucbukavu.ac.cd, zihindulabig@gmail.com

How to cite this paper: Zihindula Biguru, L. (2020) A Stochastic Strategy for Optimizing the Cost of Systematic Screening of Some Diseases in a Situation of Insufficient Equipment. *Applied Mathematics*, **11**, 1253-1274.
https://doi.org/10.4236/am.2020.1112086

Received: October 26, 2020
Accepted: December 6, 2020
Published: December 9, 2020

Abstract

In this work, we show that when there is insufficient equipment for detecting a disease whose prevalence is t% in a sub-population of size N, it is optimal to divide the N samples into n groups of size r each and then, the value $r = \left\lfloor \frac{10}{\sqrt{t}} \right\rfloor$ allows systematic screening of all N individuals by performing less than N tests (In this expression, $\lfloor x \rfloor$ represents the floor function[1] of $x \in \mathbb{R}$). Based on this result and on certain functions of the R software, we subsequently propose a probabilistic strategy capable of optimizing the screening tests under certain conditions.

Keywords

Stochastic Optimization, Binomial Distribution, Systematic Screening, Prevalence, R Software

1. Introduction

1.1. An Introductory Example

Some authors, (see [1] for example), report the problem according to which in 1945, when the American authorities were preparing to repatriate a thousand soldiers from a foreign village where they had just stayed for several months, it was learned that a dangerous disease affected 1% of the inhabitants of this village.

Systematic screening of these 1000 soldiers was necessary while various technical and logistical constraints limited the real capacity of screening tests to a

[1]Which is defined as the greatest integer less than or equal to x.

few hundred.

The teams on site took the risk of grouping the 1000 soldiers into one hundred (100) groups of ten (10) people each and mixing the samples from each group to have only 100 samples instead of 1000. For each of the 100 samples obtained, they had the names of the 10 soldiers who made it up.

Their procedure was summarized as follows:

- If the test of one of these 100 samples obtained turns out to be negative, we deduce that the 10 soldiers who constitute it are negative and in that case, a saving of nine tests will be made for this group…
- If the test of one of these 100 samples, on the other hand, turns out to be positive, they are then obliged to test each of the 10 individuals that constitute it in order to determine the health status of each of them.

In this second case, the cost of a test was wasted by carrying out eleven screenings (that of the group and that of each of the 10 soldiers who constitute it) in order to determine the health status of 10 individuals.

Various approaches make it possible to justify that for this introductory problem, the value $r=10$ soldiers per group was optimal and led to significantly less than 1000 tests to precisely determine the health situation of these 1000 individuals.

1.2. Generalization of the Problem

Starting from this very particular introductory problem, we have considered (in [2]) a more general formulation with the aim of finding a solution that is applicable to various situations when the screening capacity is clearly lower than the number of individuals test but there is a need for systematic screening.

More concretely, we consider the problem of finding the optimal value $x=r$ of the number of individuals that each group must include when we are preparing to carry out the systematic screening of N individuals for a disease whose general prevalence is t% and we also decide to divide the N samples into n groups of size r each (in this case $N=rn$).

Remark 1. *For the rest of this work, we will designate this technique described in this problem by the expression method of grouping of samples.*

2. Building an Optimal Solution

2.1. Preliminary Notions and Tools

2.1.1. Binomial Distribution

Definition 1 (Bernoulli Trial):

A Bernoulli trial is an experiment that results in two outcomes: success (S) and failure ($\overline{S}$)

The set of possible outcomes (fundamental set) of a Bernoulli trial is $\Omega=\left\{S,\overline{S}\right\}$ and it allows to define *Bernoulli random variable*.

Definition 2 (Bernoulli random variable):

Given a Bernoulli trial whose fundamental set is $\Omega=\left\{S,\overline{S}\right\}$, Bernoulli ran-

dom variable is the map $X:\Omega \to \{0,1\}$ such as $X(S)=1$ and $X(\overline{S})=0$.

Remark 2. *In a Bernoulli trial, we define the probability of success* $p=P(S)$ *and the probability of failure* $q=1-p$.

The probability mass function of a Bernoulli random variable is then given by:

$$P_X(x)=\begin{cases} p & \text{if } x=1, \\ 1-p & \text{if } x=0 \end{cases} \tag{1}$$

Definition 3 (Binomial experiment):

A binomial experiment consists of n repeated and independent Bernoulli trials where the probability p of success in each trial is known.

Definition 4 (Bernoulli random variable):

Given a binomial experiment, the binomial random variable X counts the number k of success.

Then it said that the random variable X has a binomial distribution with parameters n and p, usually written $\beta(n,p)$.

It's shown (see [3] and [4]) that the probability mass function of a binomial random X of parameters n and p is:

$$p_k = P(X=k) = C_n^k p^k (1-p)^{n-k} \quad \text{for } k=0,1,2,\cdots,n \tag{2}$$

Remark 3. *Any random variable X with a binomial distribution with parameters n and p is seen as the sum of n independent Bernoulli random variables in which the probability of success is p.*

Hence, the expected value $E(X)$ and the variance $V(X)$ of a binomial random $\beta(n,p)$ are given by:

$$E(X)=np \quad \text{and} \quad V(X)=np(1-p) \tag{3}$$

Remark 4. *For a binomial distribution* $\beta(n,p)$,

- the probability of obtaining zero success is naturally:

$$P(X=0)=C_n^0 p^0 (1-p)^n = (1-p)^n$$

- Similarly, the probability of obtaining n successes is:

$$P(X=n)=C_n^n p^n (1-p)^0 = p^n$$

Remark 5 (Linear function). It is established that if two variables X and Y are such that $Y=aX+b$ then the expected value and the variance of Y are expressed as a function of those of X by:

$$E(Y)=aE(X) \quad \text{and} \quad V(Y)=a^2V(X)$$

2.1.2. Binomial Distribution with R Software

Let us first specify that the use of the R software in this work results only from our choice and our habits.

Remark 6 (Binomial distribution with R):

As we can see in [5] and [6], R has three in-built functions to generate binomial distribution. They are described below:

- The function dbinom (k, n, p) gives the probability density distribution

$P(X=k)$ at each point $k \in \{0,1,\cdots,n\}$.

- The function pbinom (k, n, p) gives the cumulative probability $P(X \le k)$ of an event $k \in \{0,1,\cdots,n\}$.
  It is a single value representing the probability.
- The function qbinom (s, n, p) takes the probability value *s* and gives a number $k \in \{0,1,\cdots,n\}$ whose cumulative value matches the probability value *s*.

Definition 5 (Floor function). The floor function of a real number *x*, denoted $\lfloor x \rfloor$, is defined as the greatest integer less than or equal to *x*.

2.2. Solution to Introductory Example

In this section, we detail a solution of the introductory problem written in such a way that we prepare its generalization.

Let us find the number *r* of the soldiers who must constitute a group so that the total number of tests necessary to determine the health status of these 1000 soldiers is minimal.

We know in this case that, the prevalence of the disease is given and is equal to 1%.

By noting *n* the number of groups thus formed we necessarily have $nr = 1000$.

- Denote by *X* the random variable which is equal to the number of soldiers whose test will be negative in a given group. The probability *p* that the test of a soldier from this population is positive is naturally equal to the prevalence $p = 0.01$ and in this case, $q = 1 - p = 0.99$ is the probability that a any soldier be tested negative.
  As in each group there are *r* soldiers then the variable *X* is distributed according to a binomial distribution $\beta(r, 0.99)$. It then follows from the remark 4 (page 3) that the probability $P(G_n)$ that a group is negative is then:
  $$P(G_n) = P(X = r) = 0.99^r \tag{4}$$
  Similarly, $P(G_p) = 1 - P(G_n) = 1 - 0.99^r$ is the probability that a given group is positive.
- Let us denote by *Y* the random variable which is equal to the number of groups which will be tested positive among the *n*. In this case *Y* is distributed according to the binomial distribution $\beta(n, P(G_p)) = \beta(n, 1 - 0.99^r)$
  We then obtain that the expected value $E(Y)$ and the variance $V(Y)$ of the number of positive groups are respectively:
  $$E(Y) = nP(G_p) = n(1 - 0.99^r) \tag{5}$$
- Let us denote by *Z* the number of analyses necessary for the complete determination of the health status of these 1000 soldiers.
  The number *Z* of necessary tests is obtained by first performing the *n* tests corresponding to the *n* groups and then adding to them, *r* other tests for each of the *Y* positive groups.
  It follows that the random variable *Z* is written as a linear function of the variable *Y* according to the relation:

$$Z = n + rY \tag{6}$$

By relying on the remark 5 (page 3), we obtain the expression of the expected value $E(Z)$ of the number Z of the necessary tests:

$$\begin{aligned} E(Z) &= E(n + rY) \\ &= n + rE(Y) \text{ according to the remark 5} \\ &= n + r\left[n\left(1-0.99^r\right)\right] = n + rn\left(1-0.99^r\right) \\ &= n + 1000\left(1-0.99^r\right) = n + 1000\left[1-(1-0.01)^r\right] \\ &= \frac{1000}{r} + 1000\left[1-(1-0.01)^r\right] \end{aligned}$$

By considering the approximation $(1-0.01)^r \approx 1-0.01r$ we obtain:

$$\begin{aligned} E(Z) &\approx \frac{1000}{r} + 1000\left[1-(1-0.01r)\right] \\ &\approx \frac{1000}{r} + 1000(0.01r) \\ &\approx \frac{1000}{r} + 10r \end{aligned}$$

Let us denote by $N(r) = E(Z)$ the function (of variable r) which expresses the average number of tests necessary to perform to determine the health status of the 1000 soldiers.

As $N(r) = \frac{1000}{r} + 10r$ then:

$$\begin{aligned} \frac{\mathrm{d}N}{\mathrm{d}r} &= \frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1000}{r} + 10r\right) = \frac{-1000}{r^2} + 10 \\ &= \frac{10r^2 - 1000}{r^2} = \frac{10\left(r^2 - 100\right)}{r^2} \\ &= \frac{10(r-10)(r+10)}{r^2} \end{aligned}$$

As $r \in \mathbb{N}^*$, it follows from the relation $\frac{\mathrm{d}N}{\mathrm{d}r} = \frac{10(r-10)(r+10)}{r^2}$ that the signs of the derivative $\frac{\mathrm{d}N}{\mathrm{d}r}$ are the same as the signs of $(r-10)$.

And on the other hand, $\frac{\mathrm{d}N}{\mathrm{d}r} = 0$ if $r = 10$.

We then obtain the following variation table (Table 1):

Table 1. The value $r = 10$ minimizes the number of tests that are necessary for the introductory problem.

r	0		10	
$\frac{\mathrm{d}N}{\mathrm{d}r}$		−	0	+
$N(r)$	↘		min	↗

We observe indeed, that for the data of the introductory problem (N = 1000, p = 0.01), the value $r=10$ (and $n=100$) allows to minimize the average number of tests $N(r)=E(Z)=n+rE(Y)$ which then gives:

$$\begin{aligned} N_{\min} &= N(10)=\frac{1000}{10}+10E(Y) \\ &= 100+10\left[100\left(1-0.99^{1}0\right)\right] \text{ relation 5} \\ &= 100+1000\left(1-0.99^{1}0\right) \\ &\approx 196 \text{ tests.} \end{aligned}$$

Ultimately, by setting $r=10$ the number of soldiers per group, the total number of expected tests is equal to $E(Z)\approx 196$ with a standard deviation equal to:

$$\begin{aligned} \sigma_Z &= \sqrt{V(Z)}=\sqrt{V(n+rY)} \\ &= \sqrt{r^2V(Y)}=r\sigma_Y \\ &= r\sqrt{n\left(1-0.99^{r}\right)0.99^{r}} \\ &= 10\sqrt{100\left(1-0.99^{r}\right)0.99^{r}} \\ &\approx 10\times 2.940666=29.40 \end{aligned}$$

Illustration 1. note that with this value $r=10$ we are almost certain to obtain the health status of these 1000 soldiers by carrying out only less than 300 tests.

Indeed, for this value $r=10$, the probability that $Z=n+rY=100+10Y$ is less than 300 is worth:

$$P(100+10Y)\le 300=P(10Y\le 200)=P(Y\le 20)$$

Using the R syntaxes (see remark 6 page 4), we obtain for this last expression that:

```
> pbinom (20, 100, 1-0.99^10)
[1] 0.9995506
```

In other words, with the value $r=10$, there is more than 99.95% chance that the total number of tests $P(Y\le 20)=P(Z\le 300)$ required does not exceed 300 to determine the health status of these 1000 soldiers.

As announced, we will be inspired by the solution to this particular problem to build a general solution that can be transposed to various situations.

2.3. Solution to the General Problem

Consider a disease that affects t% of the population.

We have N samples corresponding to N individuals of a given sub population. In the rest of this work, we will speak indiscriminately about an individual or the sample which corresponds to him and no confusion is to be feared.

For various logistical limitations, we do not have enough equipment to perform all the N tests, but there is a need to know the health status of each of the N individuals constituting the sub population.

In the laboratory, they then decide to use the method of grouping samples as described by the remark 1 in the subsection 0.

They then to divide the N samples (it would suffice to take a part of each sample) into n groups of r individuals each and in each group we mix the r samples to run n tests. In this case, $N = rn$.

At the end of each test, two cases are possible:

- The group is negative and we deduce that all the r individuals that constitute it are negative. In this case, we save $r-1$ tests for each negative group.
- The group is positive and we are then obliged to systematically test each of the r individuals that constitute it. In this case, we waste a test because we determine the state of health of r individuals after doing $r+1$ tests (the positive group test and then the r individuals tests).

How to find the size r of each group to minimize the number of tests that it will be necessary to do to determine, in this way, the health status of all the N individuals?

According to the definition 1 (page 2), an individual's test is seen as a Bernoulli trial which leads to a success with the probability of $\frac{t}{100}$ (prevalence) and a failure with the probability $1-\frac{t}{100}$.

Denote by G_n the event *a group is negative* and G_p the event *a group is positive*.

If X denotes the random variable which is equal to the number of negative individuals in a group, the variable X is distributed according to the binomial law of the parameters r and $1-\frac{t}{100}$.

1) Let's find the probability that a group will test positive.

For a group to be negative, it is necessary that each of the r individuals that constitute it is tested negative.

So,

$$P\left(G_n\right) = P\left(X = r\right) = C_r^r \left(1-\frac{t}{100}\right)^r \left(\frac{t}{100}\right)^0 = \left(1-\frac{t}{100}\right)^r \tag{7}$$

We deduce, by using the probability of the opposite event, that the probability $P\left(G_p\right)$ for a group to be positive is equal to:

$$P\left(G_p\right) = 1 - P\left(G_n\right) = 1 - \left(1-\frac{t}{100}\right)^r \tag{8}$$

2) Law of the number of positive groups.

As we have n groups then the variable Y which is equal to the number of positive groups is a binomial distribution of the parameters n and

$$P\left(G_p\right) = 1 - \left(1-\frac{t}{100}\right)^r$$

By using the relation 3 (page 3), the expected value of number Y of positive groups and its variance $V\left(Y\right)$ are, respectively:

$$E(Y) = n \cdot P(G_p) = n\left[1 - \left(1 - \frac{t}{100}\right)^r\right] \tag{9}$$

3) The law of the total number of tests required.

Let us denote by Z the total number of tests to be carried out. It is obvious that for n tests corresponding to n groups we must add r additional tests for each of the positive groups.

It follows that the total number Z of tests is related to the number Y of positive groups by the relation:

$$Z = n + rY \tag{10}$$

According to the remark 5 (page 3), the expected value of Z is:

$$\begin{aligned} E(Z) &= E(n + rY) = n + rE(Y) \\ &= n + rn\left[1 - \left(1 - \frac{t}{100}\right)^r\right] \text{ by using the relation 9} \\ &= n + N\left[1 - \left(1 - \frac{t}{100}\right)^r\right] \text{ because } N = nr \\ &= \frac{N}{r} + N\left[1 - \left(1 - \frac{t}{100}\right)^r\right] \end{aligned}$$

Notice that in the expansion of $\text{left}\left(1 - \frac{t}{100}\text{right}\right)^r$ the terms of degrees greater than one are more and more negligible, we write:

$$\left(1 - \frac{t}{100}\right)^r \approx 1 - \left(\frac{t}{100}\right) \cdot r \tag{11}$$

In that case:

$$\begin{aligned} E(Z) &= \frac{N}{r} + N\left[1 - \left(1 - \frac{t}{100}\right)^r\right] \\ &\approx \frac{N}{r} + N\left[1 - \left(1 - \left(\frac{t}{100}\right) \cdot r\right)\right] \\ &\approx \frac{N}{r} + N \cdot \left(\frac{t}{100}\right) \cdot r \approx \frac{N}{r} + N \cdot \frac{t}{100} \cdot r \\ &\approx \frac{N}{r} + \frac{Ntr}{100} \approx N\left(\frac{1}{r} + \frac{tr}{100}\right) \end{aligned}$$

By writing $E(Z) = N\left(\frac{1}{r} + \frac{tr}{100}\right)$ where the number of individuals N and the prevalence t are known, we see that the expected value $E(Z)$ of the total number Z of tests to be performed is a function of the variable r which is the size of each group and its variation is the same as that of $f(r) = \left(\frac{1}{r} + \frac{tr}{100}\right)$ as we have:

$$E(Z) = N \cdot f(r) \tag{12}$$

4) Minimum value of $E(Z)$

$$\frac{\mathrm{d}f}{\mathrm{d}r}=\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{1}{r}+\frac{tr}{100}\right)=\frac{-1}{r^2}+\frac{t}{100}=\frac{-100+tr^2}{100r^2}$$
$$=\frac{\left(r\sqrt{t}-10\right)\left(r\sqrt{t}+10\right)}{100r^2}$$

Since $100r^2$ and $\left(r\sqrt{t}+10\right)$ are positive, then the signs of $\frac{\mathrm{d}f}{\mathrm{d}r}=\frac{\left(r\sqrt{t}-10\right)\left(r\sqrt{t}+10\right)}{100r^2}$ depend only on those of $\left(r\sqrt{t}-10\right)$ which vanishes for $r=\frac{10}{\sqrt{t}}$.

Noting that:

$$f\left(\frac{10}{\sqrt{t}}\right)=\frac{10}{\sqrt{t}}+\frac{\frac{10}{\sqrt{t}}t}{100}=\frac{\sqrt{t}}{10}+\frac{\sqrt{t}}{10}=\frac{\sqrt{t}}{5} \tag{13}$$

We deduce the following variation table (Table 2):

Ultimately, we deduce from the above that the optimal value of r is the divisor of N as close as possible to $\frac{10}{\sqrt{t}}$ where t% is the prevalence of the disease in the sub population of N individuals.

In this case, by mixing the samples from r individuals of each group, the number of tests necessary for the complete determination of the state of health of N individuals is a random variable $Z=\frac{N}{r}+r\cdot Y$ whose expected value is:

$$E(Z)=\frac{N\cdot\sqrt{t}}{5} \tag{14}$$

Then (see the definition 5 on page 4), we take $r_0=\left\lfloor\frac{10}{\sqrt{t}}\right\rfloor$ and in this case we will denote $n_0=\left\lfloor\frac{N}{r_0}\right\rfloor$ the optimal number of groups corresponding to the value r_0.

In the same way, it follows from the relation 9 that the random variable Y_0 corresponding to the number of positive groups is distributed according to a

Table 2. The optimal value of individuals per group for the general problem.

r	0		$\frac{10}{\sqrt{t}}$	
$\frac{\mathrm{d}E(Z)}{\mathrm{d}r}$		−	0	+
$E(Z)$	↘		$\frac{N\sqrt{t}}{5}$	↗

binomial law of the parameters n_0 and probability of success $p_0 = 1 - \left(1 - \frac{t}{100}\right)^{n_0}$.

To these data corresponds the optimal variable Z_0 which is equal to the total number of tests (see 10) and which is defined by:

$$Z_0 = n_0 + r_0 Y_0 \quad \text{avec} \quad E(Z_0) = E_{\min}(Z) = \frac{N \cdot \sqrt{t}}{5} \tag{15}$$

3. Towards an Optimal Strategy

In this section, we gradually establish the elements on which the optimal strategy will be based.

3.1. Preconditions

Ascertainment 1 (Low prevalence disease). The reliability of the general solution increases as the prevalence decreases.

It should be noted directly, from the expression $E(Z_0) = \frac{N\sqrt{t}}{5}$ of the expected value of the the total number Z of tests, that its value is less than the number N of individuals only if the prevalence does not exceed 25%.

On the other hand, the lower the prevalence t, the lower the number $E(Z_0) = \frac{N\sqrt{t}}{5}$ of expected tests than N.

Also, from a numerical point of view, the validity of the approximation approximation $\left(1 - \frac{t}{100}\right)^r \approx 1 - \frac{tr}{100}$ (relation 11, page 9) increases with the smallness of the prevalence t.

3.2. A Measure of the Risk of Method Failure

We propose thereafter, a succession of the syntaxes of the R software which make it possible to decide on the use of the solution to the general problem as presented in subsection 2.3.

Remark 7 (Possible values of Z_0). Note first that even when the prevalence t is very low, the value $E(Z_0) = \frac{N\sqrt{t}}{5}$ is only an expected value of the random variable $Z_0 = n_0 + r_0 Y_0$ whose set of possible values is $\{n_0, n_0 + 1, n_0 + 2, \cdots, N, N + 1, N + 2, \cdots, N + n_0\}$ where the value $n_0 = \frac{N}{r_0}$ corresponds to the first extreme case where there are no positive individuals and all the n_0 groups are then negative while the value $N + n_0$ corresponds to the extremely undesirable case where all n_0 groups are all positive.

As the objective of the strategy is to perform significantly less than N tests, the risk of failure can be measured by the probability $R = P(Z_0 \geq N)$ that the variable Z_0 takes a value which is not less than N.

So:

1) if R is substantially close to zero, we are in the case where the strategy is recommended and it will be more reliable as the risk R is close to zero.

2) if, on the other hand, R non-negligible, the strategy is contraindicated.

$$R = P\left(Z_0 \geq N\right) = P\left(\frac{N}{r_0} + r_0 Y_0 \geq N\right) = P\left(r_0 Y_0 \geq N - \frac{N}{r_0}\right)$$
$$= P\left(r_0 Y_0 \geq N\frac{\left(r_0 - 1\right)}{r_0}\right) = P\left(Y_0 \geq N\frac{\left(r_0 - 1\right)}{r_0^2}\right)$$
$$= 1 - P\left(Y_0 < N\frac{\left(r_0 - 1\right)}{r_0^2}\right) = 1 - P\left(Y_0 \leq N\frac{\left(r_0 - 1\right)}{r_0^2} - 1\right)$$

As the distribution Y_0 (number of positive groups for $r_0 = \frac{10}{\sqrt{t}}$) is binomial of the parameters $n_0 = \frac{N}{r_0}$ (the total number of groups) and $p_0 = 1 - \left(1 - \frac{t}{100}\right)^{r_0}$ (the probability for a given group to be positive) then from the relation $R = 1 - P\left(Y_0 \leq N\frac{\left(r_0 - 1\right)}{r_0^2} - 1\right)$ we can deduce a succession of the syntaxes of the software R, in accordance with the remark 6 (page 4):

First, we load the size of the sub population in the variable N and the prevalence in the variable t. We then introduce the succession of the following syntaxes:

```
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
> p0=1-((1-(t/100))^r0)
> k=N*((r0-1)/(r0^2))
> R=1-pbinom(k-1,n0,p0)
> R
```

1) For the above syntaxes;

a) $r0$ comes from the formula $r_0 = \varepsilon\left(\frac{10}{\sqrt{t}}\right)$. The trunc function calculates, for software R, the ceilling of its argument.

b) $n0$ comes from the formula $n_0 = \varepsilon\left(\frac{N}{r_0}\right)$. The trunc function guarantees the loading in the variable n_0 of an integer because r_0 may not be a divisor of N;

c) $p0$ comes from the formula $p_0 = 1 - \left(1 - \frac{t}{100}\right)^{r_0}$

d) k comes from the value $N\left(\frac{r_0 - 1}{r_0^2}\right)$ which appears in the risk expression $R = 1 - P\left(Y_0 \leq N\frac{\left(r_0 - 1\right)}{r_0^2} - 1\right)$ established in the subsection 0.

e) in the end, this last expression gives rise to the syntax R.

2) After compiling these syntaxes, the software will give the value of the risk *R*. If this value is zero, the strategy can be used.

Illustration 2 (Return to the introductory example in 1.1):

According to the data of the introductory example in 1.1 (page 1), the size of the sub population is 1000 soldiers while the prevalence is 1%.

Using the criterion gives:

```
> N=1000
> t=1
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
> p0=1-((1-(t/100))^r0)
> k=N*((r0-1)/(r0^2))
> R=1-pbinom(k-1,n0,p0)
> R
[1] 0
```

As in this case the risk of error is zero, the strategy is recommended for these data of the introductory example.

1) By calling the variables we ultimately obtain:

```
> N=1000
> t=1
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
> p0=1-((1-(t/100))^r0)
> k=N*((r0-1)/(r0^2))
> R=1-pbinom(k-1,n0,p0)
> R
[1] 0
> r0
[1] 10
> n0
[1] 100
> E=(N*sqrt(t))/5
> E
[1] 200
```

In this case, each group will actually have 10 samples, the total number of groups will be 100 and the expected value of the total number of tests to be performed is 200.

2) We can call other loaded variables, like for example p_0 and n_0 and get:

```
> p0
[1] 0.09561792
> n0
[1] 100
```

Thus, the number of positive groups $Y_0 = \beta(n_0, p_0)$ is in this introductory example, a binomial variable of the parameters $n_0 = 100$ and $p_0 = 0.09561792$ the probability that one any of the 100 groups is positive.

Illustration 3 (Contraindication). Consider the situation of a sub population of 2000 individuals for which the prevalence of a disease is 30%.

By measuring the risk of strategy failure we obtain:

```
> N=2000
> t=30
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
> p0=1-((1-(t/100))^r0)
> k=N*((r0-1)/(r0^2))
> R=1-pbinom(k-1,n0,p0)
> R
[1] 1
>
```

So $R = 1$ means that at 100% the method will fail.

3.3. Effectiveness of the Method

By exploiting the illustration made above on various values of the prevalence, we note that this method exposed in (3) can only be used when the prevalence does not exceed t = 25%, a case in which it is almost impossible for the method to lead to more tests than the N samples.

It should be noted, from the illustrations above that the risk R that the method leads to more tests than samples, almost coincides with the function $f :]0,100[\to [0,1]$ and defined by:

$$R(t) \equiv f(t) = \begin{cases} f(t) \approx 0 & \text{si } t \leq 25\% \\ f(t) \approx 1 & \text{si } t > 25\% \end{cases} \tag{16}$$

From the syntax t = seq(1,30,0.1) of software R we take, for the introductory example, all the values of the prevalence ranging from 1% to 30% with an increment of 0.1% and by executing the syntax below we obtain the graphical representation of the variation of the risk of failure R according to the prevalence:

```
> t=seq(1,30,0.1)
> r0=trunc(10/sqrt(t))
> N=1000
> n0=trunc(N/r0)
> p0=1-((1-(t/100))^r0)
> k=N*((r0-1)/(r0^2))
> R=1-pbinom(k-1,n0,p0)
> plot(t,R)
```

The compilation of these commands gives graphical representation (**Figure 1**).

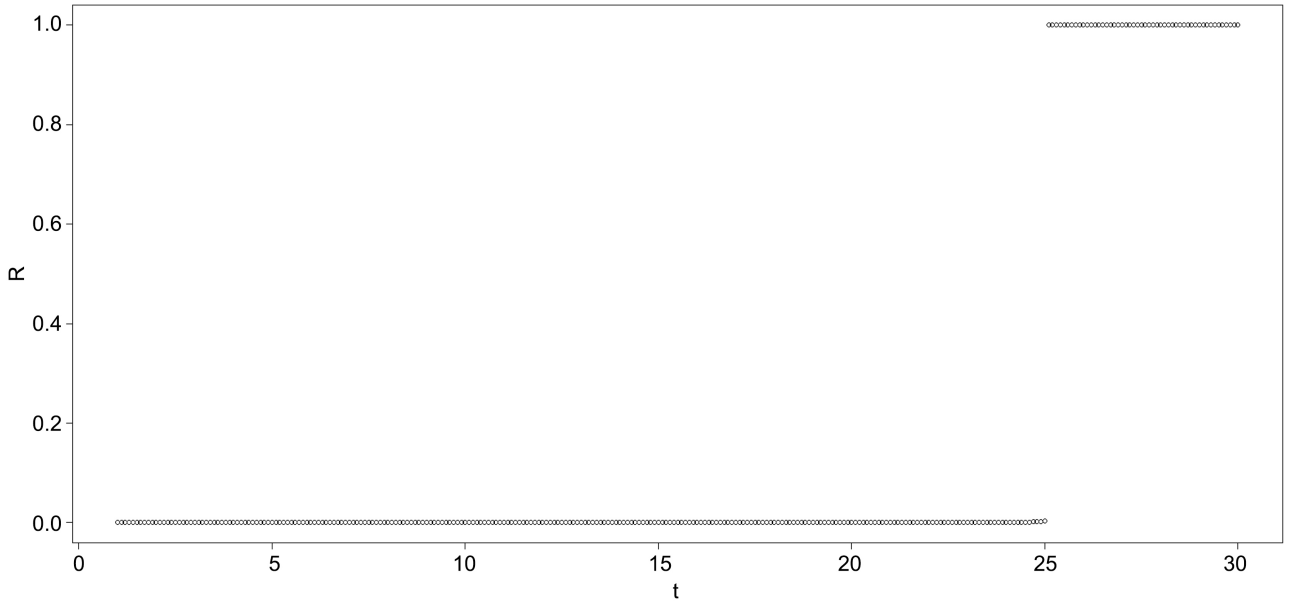

Figure 1. Represents the probability R of the failure of the method as a function of the prevalence t which is expressed as a percentage.

Considering the fact that the manipulations (mixing of samples, storage, …) Required by the method present a certain cost, it is not sufficient to establish that for small values of the prevalence of less than 25%, the method will be efficient.

Depending on certain specificities of the test and the nature of the samples, the method could lead to slightly less than N tests without the cost of the difference in the tests compensating for the cost of handling and using the method.

There then arises the need for another indicator of the effectiveness of this method.

Remark 8 (Effectiveness of the strategy). Beyond the risk R established as almost zero for $t < 25\%$, we consider that the method is effective when it leads to a number of tests well below half $N/2$ of the total number samples.

In order to develop syntaxes R making it possible to decide on the efficiency E of the proposed strategy, within the meaning of the remark above, let's calculate $E = P\left(Z_0 < N/2\right)$, the probability that the total number of tests is less than half of the number of samples.

$$\begin{aligned} E &= P\left(Z_0 < \frac{N}{2}\right) = P\left(\frac{N}{r_0} + r_0 Y_0 < \frac{N}{2}\right) \\ &= P\left(r_0 Y_0 < \frac{r_0 N_0 - 2N}{2r_0}\right) = P\left(r_0 Y_0 < \frac{N\left(r_0 - 2\right)}{2r_0}\right) \\ &= P\left(Y_0 < \frac{N\left(r_0 - 2\right)}{2r_0^2}\right) = P\left(Y_0 \le \frac{N\left(r_0 - 2\right)}{2r_0^2} - 1\right) \end{aligned}$$

From the relation $E = P\left(Y_0 \le \frac{N\left(r_0 - 2\right)}{2r_0^2} - 1\right)$ where Y_0 is a binomial distribution of the parameters n_0 and p_0 we deduce, from the remark 6, (page 4),

that this efficiency can be calculated directly by the syntax pbinom(l-1, n0, p0) with $l=\frac{N(r_0-2)}{2r_0^2}$ on condition that all the values prior to loading l are already saved in the R environment.

Illustration 4 (Conditions of effectiveness of the introductory example). Returning to the introductory example in 0, the prevalence is equal to $t=1\%$ and $N=1000$, we obtain the efficiency using the following commands:

```
> t=1
> N=1000
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
> p0=1-((1-(t/100))^r0)
> l=N*((r0-2)/(2*r0^2))
> E=pbinom(l-1,n0,p0)
> E
[1] 1
```

As for this introductory example (see 1.1), $E=1$, we deduce that by taking $r_0=10$ and $n_0=100$, the probability of performing less than 500 tests and appreciably close to 1. The method is therefore effective for this case.

Illustration 5 (Effectiveness of the method as a function of prevalence). With the command t = seq(1,25,0.1) we load all the prevalences between 1% and 25% and using the commands below we call the graphical representation of the efficiency $E(t)=P\left(Z_0<\frac{N}{2}\right)$ as a function of the prevalence t.

```
> t=seq(1,25,0.1)
> N=1000
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
> p0=1-((1-(t/100))^r0)
> l=N*((r0-2)/(2*r0^2))
> E=pbinom(l-1,n0,p0)
> plot(t,E)
```

The execution of these commands gives the Figure 2.

By extracting some values from the function $E(t)$ (the command E [i] on R gives the ith component of the vector E) we notice that:

Table 3 confirms that for the introductory problem, the efficiency of the method is almost certain as long as the prevalence t does not exceed 5% and it decreases as t increases.

Table 3. Efficiency of the method for the introductory problem.

t	1%	2%	3%	4%	5%	6%	7%	8%	…
$E(t)$	1	1	1	0.9999471	0.9915099	0.8465961	0.06821181	0.004509621	…

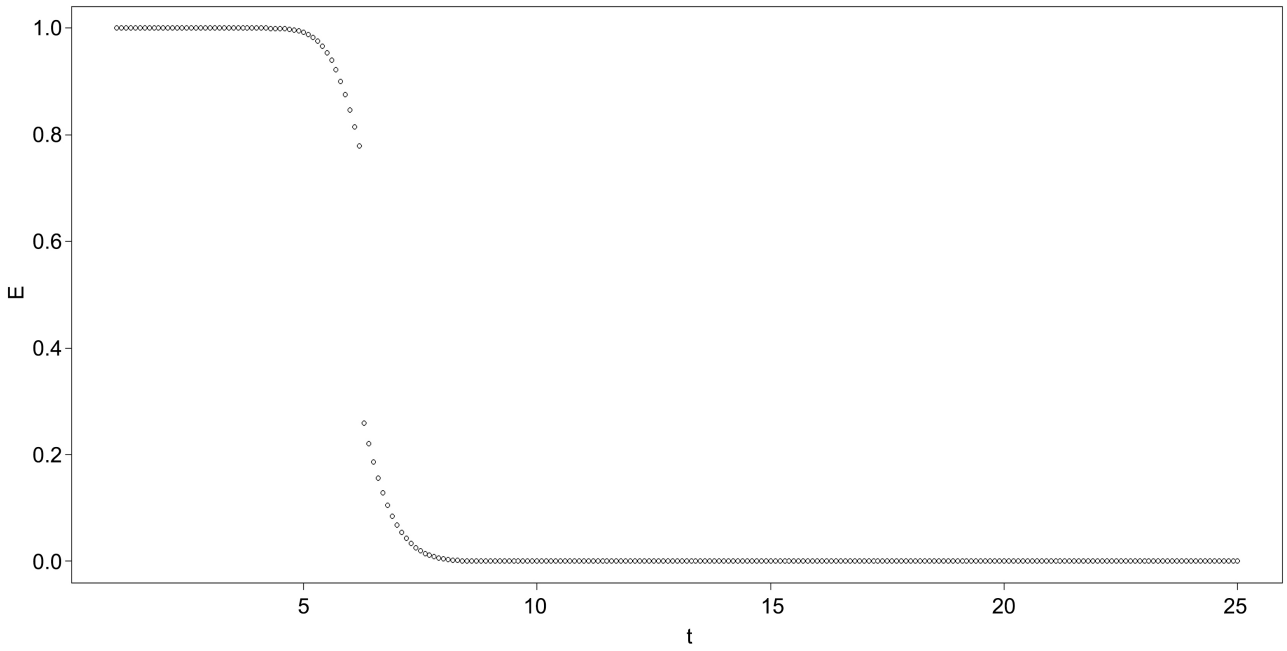

Figure 2. Represents the efficiency E of the proposed method as a function of the prevalence t which is expressed as a percentage.

From what precedes we deduce the following statement:

Ascertainment 2 (Conclusion of the introductory example): For the $N = 1000$ individuals of the introductory example, the method could be applied with guarantee of determining the health status of all 1000 individuals by performing less than 500 tests provided that the prevalence t does not exceed 5%.

4. Optimal Strategy

4.1. Specific Objectives of the Strategy

When using the **method of grouping samples** (see remark 1, page 2), the objective of our strategy is to provide test practitioners with a concrete approach that should enable them, in various circumstances, to:

- determine the optimal size of each group as done on 2.3,
- measure the risk of failure of this method as done on 3.2,
- measure the effectiveness of the method and as done on 3.3,
- find the limit value of the number of tests that the use of the method will not exceed.

The combination of these informations will allow to decide whether or not to use the method of grouping the samples.

At this level it is, in fact, possible to compare the cost entailed by the use of the method of grouping samples with that of the $N - Z_{\max}$ tests that the method saves.

4.2. Foundations of the Strategy

4.2.1. Main ingredients

- In subsection 2.3, we have established that by taking $r = \frac{10}{\sqrt{t}}$, the method

consisting of dividing the N samples into n groups of r individuals each makes it possible to minimize the expected value $E(Z)$ of the random variable Z which is equal to the total number of tests necessary to determine the health status of all the N individuals.

- In 3.1 and 3.2 we have specified that as long as the prevalence t of the disease does not exceed 25% the risk R of failure of this method is almost zero but that low prevalence values are most recommended.
- In 3.3, we denote by $E(t)$ the probability that the method leads to less than $\frac{N}{2}$ tests, a probability that we have assimilated to **a indicator of efficiency** of the method.

It is on the basis of these results that we identify the optimal strategy of using the **method of grouping samples** as described in the subsection 1.2 (page 2).

At this stage only the calculation of $Z_{\max}$ is missing.

4.2.2. Determination of $Z_{\max}$

Let us calculate the limit value $Z_{\max}$ of the number of tests that we will not exceed.

To calculate the limit value $Z_{\max}$ that the random variable Z (the total number of tests) will not exceed, it is necessary to first fix the error level η (or the reliability $1-\eta$) with which the limit value z will be determined.

In practice, the risk of error η must be so low (we can take for example $\eta \approx \frac{1}{10000}$) that we are almost certain that in fact, $Z \le Z_{\max}$.

It's important to note that in subsection 0.0.2 we deduced from the relation 10 that , for the optimal value r_0, the total number of tests Z_0 is written in the form $Z_0 = n_0 + r_0 Y_0$ where Y_0 is is the number of positive groups and Y_0 is a binomial distribution $\beta(n_0, p_0)$.

After having fixed a very low value of the error η we have:

$$\begin{aligned} &P(Z_0 \le Z_{\max}) = 1-\eta \Rightarrow P(n_0 + r_0 Y_0 \le Z_{\max}) = 1-\eta \\ &\Rightarrow P\left(Y_0 \le \frac{Z_{\max} - n_0}{r_0}\right) = 1-\eta \end{aligned} \tag{17}$$

By exploiting the relation 17:

$$l = \frac{Z_{\max} - n_0}{r_0} \Leftrightarrow Z_{\max} = n_0 + r_0 l \tag{18}$$

we obtain that:

$$P(Y_0 \le l) = 1-\eta \text{ where } Y_0 \equiv \beta(n_0, p_0) \text{ is a binomial distribution.} \tag{19}$$

According to the third in-built function of the R software (see 2.1.2) we deduce from relation 19 that:

$$l = qbinom(1-\eta, n_0, p_0) \text{ and then from 18 we have:} \tag{20}$$

$$Z_{\max} = n_0 + r_0 l = n_0 + r_0 * qbinom(1-\eta, n_0, p_0) \tag{21}$$

Illustration 6 (Maximum number of tests of the introductory example):

The introductory example presented in 0 we have a prevalence $t=1\%$ for a sub population of size $N=1000$.

As we established in subsection 3.2, the optimal parameters of the *sample grouping method* are loaded into the R software as follows:

```
> N=1000
> t=1
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
> p0=1-((1-(t/100))^r0)
```

If we fix, for example, a risk of error $\eta=\frac{1}{10000}$, we can use the syntax of relation 21:

```
> eta=1/10000
> Zmax=n0+r0*(qbinom(1-eta, n0, p0))
```

By executing all of these syntaxes in the R software environment we ultimately obtain:

```
> N=1000
> t=1
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
> p0=1-((1-(t/100))^r0)
> eta=1/10000
> Zmax=n0+r0*(qbinom(1-eta , n0, p0))
> Zmax
[1] 320
```

We interpret this result by asserting that for the introductory example, there are 9999 out of 10,000 chances, *i.e.* 99.99%, that using the parameter $r_0=10$ (and in this case $n=100$) will lead to a total number of tests that does not exceed 320 for all the 1000 samples.

4.3. Summary of the Optimal Strategy

4.3.1. Descriptive Steps of the Optimal Strategy

Given the need to perform a systematic screening of a disease with a prevalence of *t*% within a sub population of size *N*, the application of the optimal strategy to the method of grouping samples as described by the remark 1 (page 2) consists of the following four steps:

Step 1 (Determination of method parameters). At this level we determine the size r_0 of each group on which depend the parameters of the random variable Y_0 which is equal to the number of positive groups among all the n_0 groups.

1) Y_0 is a binomial random variables of parameters n_0 and $p_0=1-\left(1-\frac{t}{100}\right)^{r_0}$;

2) The random variable Z_0 which is equal to the total number of tests to be performed is related to Y_0 by the relation: $Z_0 = n_0 + r_0 Y_0$

Step 2 (Measurement of the risk of method failure). $R = P(Z_0 > N)$ R is the probability that this method leads to more tests than samples.

- If $R \approx 0$ then the method can be used.
- If $R \neq 0$ then the method cannot be used.

Step 3 (A measure of the effectiveness of the method). $E = P\left(Z_0 \leq \frac{N}{2}\right)$ E is the probability that the total number of tests carried out by the method does not exceed half of the total number of samples.

- If $R \approx 0$ and $E \approx 1$ then the use of the method is strongly recommended.
- If $R \approx 0$ and $E \neq 1$ then the final decision will be taken by comparing the total cost of using the method of grouping of samples with the cost of $N - Z_{\max}$ tests that will be saved by the method.

At this level it only lacks the calculation of the maximum number $Z_{\max}$ of tests that the method will not exceed.

Step 4 (determination of $Z_{\max}$). As shown in 4.2.2, a syntax of R software makes it possible to determine $Z_{\max}$ with a desired precision $1-\eta$ (or with a risk η of error).

4.3.2. Expression of the Strategy in the R Software Environment

1) To perform the first step we use the following syntax:

```
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
```

2) To perform the second step we use the following syntax (see 3.2):

```
> p0=1-((1-(t/100))^r0)
> k=N*((r0-1)/(r0^2))
> R=1-pbinom(k-1,n0,p0)
> R
```

3) To carry out the third step we use the following syntax (see 3.3):

```
> l=N*((r0-2)/(2*r0^2))
> E=pbinom(l-1,n0,p0)
> E
```

4) To carry out the fourth step we use the following syntax:

If we fix a risk of error η, we can use the syntax of relation 21:

```
> eta=v
> Zmax=n0+r0*(qbinom(1-eta, n0, p0))
> Zmax
```

4.3.3. Examples

Example 1. Let consider a disease with a prevalence of 5% in a subpopulation of 5000 individuals.

1) Using the first step we get:

```
> N=5000
> t=5
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
> r0
[1] 4
> n0
[1] 1250
>
```

It is therefore optimal to mix 4 samples for each of the 1250 groups.

2) Using the second step we get:

```
> p0=1-((1-(t/100))^r0)
> k=N*((r0-1)/(r0^2))
> R=1-pbinom(k-1,n0,p0)
> R
[1] 0
>
```

As $R = 0$ then the method can be used.

3) Using the third step we get:

```
> l=N*((r0-2)/(2*r0^2))
> E=pbinom(l-1,n0,p0)
> E
[1] 1
>
```

As $E = 1$, it is certain that the method will lead to a number of tests which does not exceed half of the samples. The method is strongly recommended.

Using the fourth step we have by choosing, for example $\eta = \frac{1}{10000}$:

```
> eta=1/10000
> Zmax=n0+r0*(qbinom(1-eta, n0, p0))
> Zmax
[1] 2386
>
```

We conclude this example by asserting that the value $r_0 = 4$ is optimal.

- As $E = 1$ it is certain that the method will lead to a number of tests which does not exceed 2500.
- As $Z_{\max} = 2386$ for $\eta = \frac{1}{10000}$, we affirm, with a probability of 0.9999, that the value $r_0 = 4$ will allow systematic screening of these 5000 individuals by performing a number of tests which does not exceed 2386.

and in this case there is a probability of 99.99% that the total number of tests

does not exceed 434.

Example 2. Consider a disease with a prevalence of 10% in a sub population of 600 individuals.

1) Using the first step we get:

```
> N=600
> r0=trunc(10/sqrt(t))
> n0=trunc(N/r0)
> r0
[1] 3
> n0
[1] 200
```

It is therefore optimal to mix three samples for each of the 200 groups.

2) Using the second step we get:

```
> p0=1-((1-(t/100))^r0)
> k=N*((r0-1)/(r0^2))
> R=1-pbinom(k-1,n0,p0)
> R
[1] 0
>
```

As $R = 0$ then the method can be used.

3) Using the third step we get:

```
> l=N*((r0-2)/(2*r0^2))
> E=pbinom(l-1,n0,p0)
> E
[1] 1
>
```

In this case ($E \approx 0$), even if the risk of failure is zero but it is almost impossible that the method leads us to less than 300 tests.

4) Using the fourth step we have by choosing, for example $\eta = \frac{1}{10000}$:

```
> eta=1/10000
> Zmax=n0+r0*(qbinom(1-eta, n0, p0))
> Zmax
[1] 434
>
```

We conclude this example by asserting that the value $r_0 = 3$ is optimal and in this case there is a probability of 99.99% that the total number of tests does not exceed 434.

5. Conclusions and Perspectives

The strategy proposed in this work responds to the need to carry out systematic

screening when logistical resources are not sufficient.

The conditions of use (low prevalence) of this strategy are frequent in practice. In this period when humanity is facing the COVID-19 pandemic, there is almost everywhere a lack of screening equipment.

Unfortunately, we cannot yet recommend our strategy in the case of covid-19 because, on the one hand, the prevalence is generally not known and, on the other hand, the screening tests and the nature of the samples considered must still progress.

In terms of perspectives, we will focus the rest of this research on a double project:

- From a mathematical point of view, we will try to integrate the estimation of the prevalence in this strategy in order to be able to apply it to diseases whose prevalence is not known.
- Regarding applications, we would like to increase the number of exchanges with epidemiology researchers in order to determine the diseases and tests for which the type of sample allows this strategy to be applied.

Acknowledgements

This work was born from studious discussions that I led with some colleagues from the Faculty of Sciences of the Catholic University of Bukavu (UCB) as well as those of the Faculty of Sciences and Applied Sciences of the Official University of Bukavu (UOB). This article gives me the best opportunity to express my gratitude to them.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Audry, D. and Maths Terminale, S. (2002) Enseignement Obligatoire. Centre National d'Enseignement à Distance (CNED), Institut de Rennes, Rennes.

[2] Zihindula Biguru, L. (2010) Une optimisation stochastique du cout de dépistage de certaines maladies rares dans une population donnée. Annale des Sciences de l'Université officielle de Bukavu, Vol. 2, Bukavu.

[3] Florescu, I. and Tudor, C. (2013) Handbook of Probability. Wiley. https://doi.org/10.1002/9781118593103

[4] Ross, S. (2007) A First Course in Probability. Academic Press.

[5] Core Team, R. (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

[6] Dolores Ugarte, M., Militino, A. and Arnholt, A. (2015) Probability and Statistics with R. 2nd Edition, CRC Press. https://doi.org/10.1201/b18682