Estimation of prevalence in rare disease using pooled samples

J P Martins[1,2], R Santos[1,2] and M Felgueiras[1,2,3]

[1] School of Technology and Management, Polytechnic Institute of Leiria, Campus 2, Morro do Lena Alto do Vieiro, Apartado 4163, 2411-901 Leiria, Portugal
[2] CEAUL, Centre of Statistics and its Applications
[3] CARME, Polytechnic Institute of Leiria

E-mail: jpmartins@ipleiria.pt

Abstract. The use of pooled samples for screening infected individuals is a known procedure to reduce costs. In an estimation problem, the aim is only to determine how many individuals are infected instead of determining who is infected (classification problem). In that setting, our goal was to compare the performance of using one or two-dimensional arrays. The best performance was established according to one of the following criteria: minimizing the number of individuals or the number of tests required to attain a certain estimate accuracy. It is observed that when we want to minimize the number of individuals used, the two-dimensional procedures have a little advantage over the one-dimensional procedures. However, when the major concern is the cost, the one-dimensional procedures clearly outperform the two-dimensional procedures.

1. Introduction

The problem of using pooled samples for screening infected individuals among a population of size N dates back from 1943 Dorfman's seminal work [1]. Dorfman's two stage procedure consists in homogeneously mixing the samples from n individuals for batched testing. A positive result is inconclusive in the way that both the number of infected samples and who is infected is unknown. This unclear result leads to a second stage where individual tests are performed. On the other hand, in a perfect test, a pooled negative result clearly means that none of the individuals is infected.

The research in this field may be split in two according to its goal. In one case, the aim is to identify *who* is actually infected (classification problem). The other is to determine *how many* are infected (estimation problem). We will be restricted to this last goal. Moreover, we will only consider low values of the prevalence rate (lower than 5%, and therefore we will be working with a rare event).

The aim of using pooled samples is to save money [2] as the number of tests performed may be lower than the number of tests required by individual testing [3]. As we just want to determine how many individuals are infected, the performance of individual testing is optional whereas it is mandatory for positive pooled samples at some stage when dealing with a classification problem.

This freedom of dismissing individual testing allows us to think in complex schemes. Robotic pooling allows us, in practice, to effectively and accurately take advantage of more complex mixing procedures since the costs of mixing samples are usually negligible [4]. Pooled samples

may or may not have individuals in common. They can be retested despite their limited benefits [5] or divided into subsamples for further testing (hierarchical models).

Mixing a large number of individuals into the same pool is a reasonable possibility although it is important to have in mind the dilution effect [3]. The accuracy of a test are usually assessed by its sensitivity and specificity. They define the probability of correctly classifying an infected and a non-infected sample, respectively. In practice, these probabilities may decrease with the pool size.

In this setting, three questions arise when one-stage procedures are considered. The first is to decide what procedure to use, that is, what kind of pools should be used? Then, two further questions arise that may be gathered in just one question: How many pools and individuals within the same pool should be used?

The answer of this questions is closely related to an underlying issue: How can we compare different methodologies? The answer to this question does not depend only in the quality of the results but may also depend on the samples availability or the cost of the procedure. To assess the quality of the results there are a wide variety of measures. We chose the root mean square error (RMSE) for this purpose. As previously stated, the cost of a methodology largely depends on the number of tests performed. Hence, the cost mainly increases with the number of tests performed. However, in other situations money is not the main issue. The individual samples may be difficult to get. Thus, we may want to reduce the number of individuals used.

In our work, we found that two-dimensional arrays are the best option when we want to use as few individuals as possible whereas when the cost of the test is the most relevant issue one dimensional arrays are recommended.

The outline of this work is as follows. Section 2 introduces some of the most common pooling samples procedures as well as describe some restraints that lead to a particular choice of a procedure. Next, in Section 3 the results of a simulation study are presented and discussed. Finally, in Section 4 the final conclusions are presented.

2. One stage procedures in a classification problem

Let $Y_1, \cdots, Y_N$ be a sample drawn from a population where each individual Y_i is a dummy variable with values 1 and 0 with probabilities p and $1-p$ respectively. The parameter p stands for the prevalence rate. Performing individual tests is probably the simplest way to estimate p, although it can be quite inefficient [6, 7, 8]. Hence, the use of pooled samples may be an option. In what follows, individual status within a pooled sample is assumed to be independent and no dilution effect is considered. Previously described Dorfman's methodology comprehends two stages (a pooled test and if the outcome is positive subsequent individual tests). In this work, we deal only with one stage procedures in the way that the results obtained in any pooled sample do not lead to further tests.

In this setting, it is possible to consider the use of non-overlapping pools called one dimensional arrays. If individual samples can be present in more than one pool, the most popular design are square arrays (a two-dimensional design). The use of more than two dimensions has proved to be not advised in most of the situations [9, 10].

2.1. One dimensional procedures

A one-stage and one-dimensional procedure is probably the simplest and easiest pooling method. The design of the procedure depends only in setting the pool size n. For this reason, we will refer to it as procedure $O(n)$.

For simplicity, admit that n is a divisor of the sample size N (otherwise, one would have $\left[\frac{N}{n}\right]$ groups with n individuals and one group with $N - \left[\frac{N}{n}\right] \times n$ individuals, where $[x]$ stands for the highest integer lower than x). Then, it is required to perform $T_N = \frac{N}{n}$ tests. Clearly, an infected

pooled sample is found with probability $\pi_n = 1-(1-p)^n$. Hence, the total number of infected pooled samples is described by a binomial random variable $I \frown \text{Bin}(T_N, \pi_n)$, where T_N is the trials number and π_n the success probability. Thus, the maximum likelihood (ML) estimator of π_n is given by

$$\widehat{\pi_n} = \frac{I}{T_N}. \tag{1}$$

As p and π_n are directly related, the ML estimator of p is given by

$$\widehat{p} = 1 - \left(1 - \frac{I}{T_N}\right)^{1/n}. \tag{2}$$

For $n = 1$, $\hat{p} = 1 - \left(1 - \frac{I}{T_N}\right) = \frac{I}{T_N}$ is an unbiased estimator of p. For $n > 1$, the estimator is positively biased. Expressions for the expected value and variance of the estimator can be found in [11].

As screening errors may occur, the above binomial model is, in practice, unrealistic. The probability of observing a pooled positive result is $\varphi_s + (1-\varphi_s-\varphi_e)(1-p)^n$ as stated in [12]. Therefore, a ML estimator of p is

$$\widehat{p} = 1 - \left(\frac{\varphi_s - p^+}{\varphi_s + \varphi_e - 1}\right)^{1/n}. \tag{3}$$

where $p^+ = \frac{P}{T_N}$ is the proportion of positive results The estimator only assumes meaningful values if

$$1 - \varphi_e \leq p^+ \leq \varphi_s. \tag{4}$$

2.2. Two-dimensional procedures

In this setting, square arrays are probably the most used pooling procedure that uses overlapping samples. Suppose we have n^2 individuals placed in a $n \times n$ matrix. All individuals within the same row and within the same column are gathered for batched testing.

Even if a perfect test is used, the outcomes are not readily clear. Having r positive rows and c positive columns only means that there are a maximum of $r \times c$ infected individuals.

If the test is not perfect, ambiguities may arise. For instance, in a square array procedure one may have both a positive row and all columns testing negative. Clearly, the proportion of infected individuals is not the best option in this setting. To avoid discarding ambiguous results, [13, 14] present a first algorithm to overcome this issue, which was improved in [12]. This procedure will be denoted as $T(n)$.

The computation script developed involves some simulation performance. The main idea is to use simulation to find an estimate of the value p_0 that minimizes the function

$$Dif(p_0|O) = \sum_{i,j} \left(O(i,j) - s \times P_{p_0}(i,j)\right)^2, \tag{5}$$

where matrix O encodes the number of square arrays having $i-1$ $(j-1)$ positive rows (columns) for $i = 1, 2, \cdots, r+1$ $(j = 1, 2, \cdots, c+1)$ and s is the total number of two-dimensional arrays (i.e., $s = \sum_{i,j} O(i,j)$). The matrix P_{p_0} encodes the estimates of the probability for each entry of matrix O given a prevalence rate p_0.

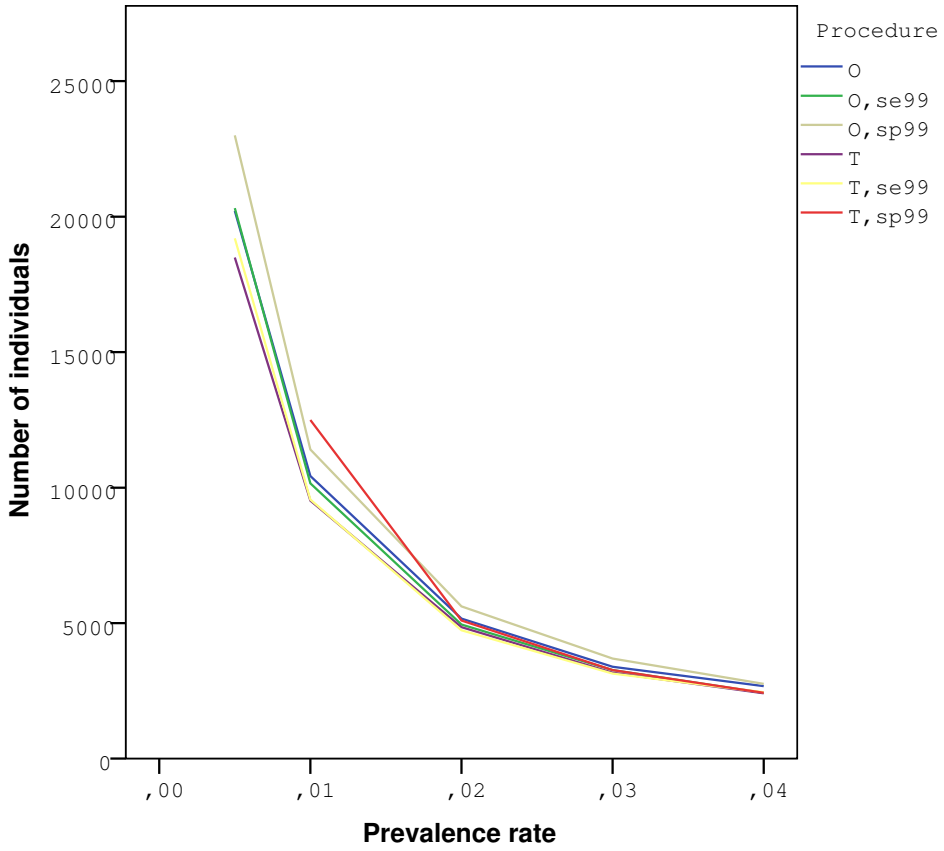

Figure 1. Minimum number of individuals required to attain the RMSE goal for several prevalence rates (O – one dimensional, T – two dimensional, the tests are assumed perfect unless when identified with: se99 – sensitivity equal to 0.99 or sp99 – specificity equal to 0.99)

3. Results

To assess the performance of the previous two procedures a simulation study was conducted. The underlying prevalence rates used were $0.005, 0.01, 0.02, 0.03$ and 0.04. The target RMSE was a RMSE lower than 10% of the true prevalence rate. The experimental test was assumed with sensitivity and specificity ranging from 0.95,0.99 to 1. Both measures are not necessarily equal.

No more than 50 individuals were considered within the same pool due to the dilution problem. In both procedures, the maximum number of pools was set equal to 100.

For the two-dimensional arrays, 2000 replicas of $k \times T\,(n)$ arrays were simulated for increments of 0.0005 of the prevalence rate where $k = 1, \cdots, 120$ is the number of arrays.

Figure 1 summarizes some of the results when the goal is to minimize the number of individuals involved whereas Figure 2 displays the results obtained when the cost of the procedure is the main concern.

Concerning the Figure 1, the number of individuals decreases with the true prevalence rate. The array size decreases, in general, with the true prevalence rate (results not shown). Decreasing the test specificity has a greater impact in the number of individuals compared to decreasing the test sensitivity. Moreover, the RMSE goal was not attained when the test specificity was equal to 0.99 for the lowest prevalence rate value. A perfect test is more important when the prevalence rate is low although the perfect test behaves worse than the test with perfect specificity and

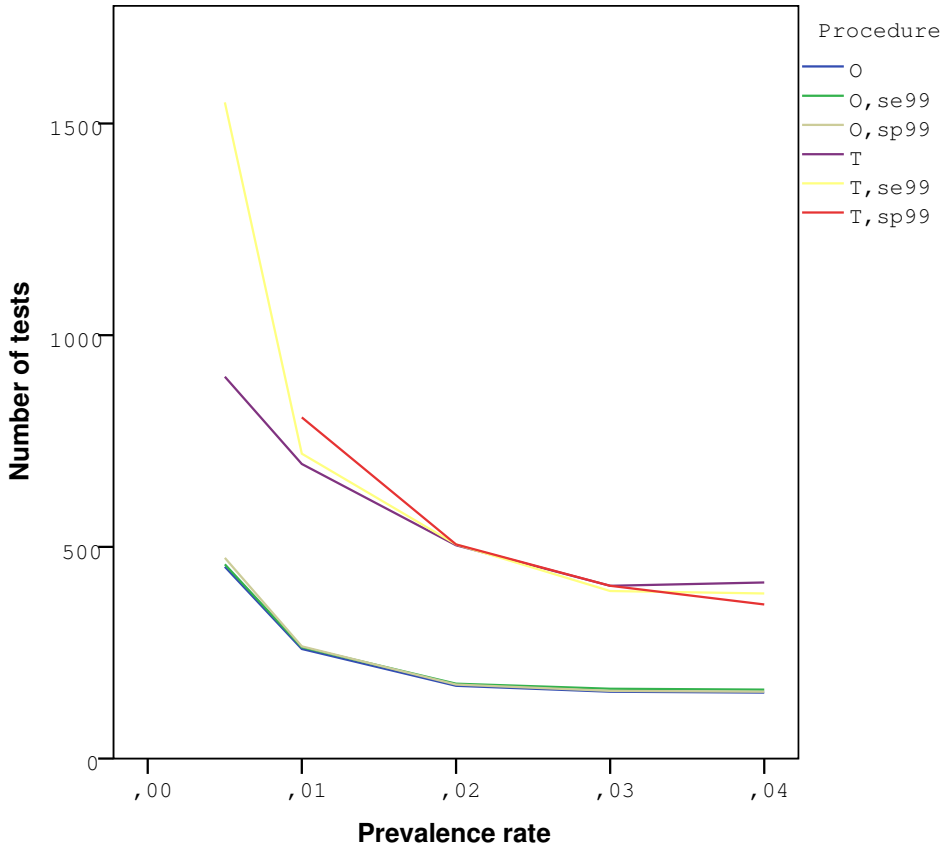

Figure 2. Minimum number of tests required to attain the RMSE goal for several prevalence rates (O – one dimensional, T – two dimensional, the tests are assumed perfect unless when identified with: se99 – sensitivity equal to 0.99 or sp99 – specificity equal to 0.99)

sensitivity equal to 0.99. This may be caused by the low number of positive results. The two-dimensional procedures outperform the one-dimensional procedures. However, the results are not very different.

Observing Figure 2 the different procedures can have quite different performances.

The number of tests decreases with the true prevalence rate. The array sizes also start to decrease at some point when the prevalence rate increases (results not shown). For one case, two-dimensional procedure with a test whose specificity was 0.99 did not allow to obtain the desired RMSE. When the prevalence rate is not higher than 1%, decreasing the test specificity induces the performance of a higher number of tests when compared to a similar decrease of the test sensitivity. The opposite effect is verified for the other values of p. The perfect tests produced once again better results especially for the lower values of the prevalence rate. The one-dimensional procedures outperform the two-dimensional procedures with significant gains.

4. Conclusion

The conjecture presented in [14] which suggests that a high number of dimensions in estimation problems may not be very useful was at some extent observed in this work. However, [9] found some advantages of using three-dimensional arrays in classification problems. The performance of two-dimensional arrays when it is important to minimize the number of individuals used is recommended. However, the use of one-dimensional array procedures only involves a small

increase of the number of individuals.

On the other hand, when the main concern is to reduce the number of tests performed, the use of one-dimensional arrays is strongly recommended as the savings are significantly larger.

Acknowledgments
Funded by FCT – Fundação para a Ciência e Tecnologia, Portugal, through the project UIDB/00006/2020.

References

[1] Dorfman R 1943 The detection of defective members in large populations *Ann. Math. Stat.* **14(4)** pp 436-40
[2] Gastwirth J L and Johnson W O 1994 Screening with Cost-Effective Quality Control: Potential Applications to HIV and Drug Testing *JASA* **89(427)** pp 972-81
[3] Santos R, Pestana D and Martins J P 2013 Extensions of the Dorfmans theory *Recent Developments in Modeling and Applications in Statistics, Studies in Theoretical and Applied Statistics* ed Oliveira P E *et al* (Berlin: Springer-Verlag) pp 179-89
[4] Liu S C, Chiang K S, Lin C H, Chung W C, Lin S H and Yang L C 2011 Cost analysis in choosing group size when group testing for Potato virus Y in the presence of classification errors *Ann. Appl. Biol.* **159(3)** pp 491–502
[5] Chen C L and Swallow W H 1990 Using group testing to estimate a proportion, and to test the binomial model *Biometrics* **46** pp 1035-46
[6] Garner F C, Stapanian M A, Yfantis E A and Williams L R 1989 Probability estimation with sample compositing techniques *J. Off. Stat.* **5** pp 365-74
[7] Loyer M W 1983 Bad probability, good statistics, and group testing for binomial estimation *Am. Stat.* **37** pp 57-9
[8] Sobel M and R M Elashoff 1975 Group testing with a new goal, estimation *Biometrika* **62** pp 181-93
[9] Kim H and Hudgens M 2009 Three-Dimensional Array-Based Group Testing Algorithms *Biometrics* **65(3)** pp 903-10
[10] Martins J P, Felgueiras M and Santos R 2015 Three-Dimensional Array-Based Group Testing Algorithms With One-Stage *AIP Conf. Proc.* **1702(1)** 030004
[11] Hung M and Swallow W H 1999 Robustness of group testing in the estimation of proportions *Biometrics* **55** pp 231-7
[12] Martins J P, Felgueiras M and Santos R 2017 Estimation through array-base group tests *RevStat* **15(4)** pp 487-500
[13] Martins J P, Felgueiras M and Santos R 2014 Maximum Likelihood Estimation in Pooled Sample Tests *AIP Conf. Proc.* **1618(1)** pp 543-6
[14] Martins J P, Santos R and Felgueiras M 2015 A Maximum Likelihood Estimator for the Prevalence Rate Using Pooled Sample Tests *Theory and Practice of Risk Assessment, Springer Proceedings in Mathematics & Statistics* **136** ed Kitsos C P *et al.* (Berlin: Springer) pp 99-110