Missing two key parameters in the pooled testing strategy

Xuhua Xia (xxia@uOttawa.ca)
University of Ottawa https://orcid.org/0000-0002-3092-7566

Biological Sciences - Article

Keywords: pooled testing strategy, infection prevalence, variance

Posted Date: October 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1028619/v1

Missing two key parameters in the pooled testing strategy

Xuhua Xia[a]*

[a]*Department of Biology and Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Canada*

*corresponding author

The pooled testing strategy [1] misses two key parameters, the infection prevalence *p* and its variance mentioned many times in the paper as the key determinants of any pooled testing strategy. For illustrating their methods, the authors used *p* from other studies that employed individual tests. It turned that that no statistical estimators for *p* and its variance have ever been derived for testing data of pooled samples since the first formulation of testing strategies based on pooled sampled in 1943 [2]. Here I derive the maximum likelihood estimators for *p* and its variance based on tests of pooled samples. This should result in significant saving in time, resource, and costs.

Define *N* as the number of individuals sampled from a population, and *n* as the pool size, i.e., the number of samples pooled together in a single test. If $N = 1{,}000{,}000; n = 100$, then 10,000 tests ($= N/n$) need to be done. The maximum *n* value is limited by the sensitivity of the test. In the case of COVID-19 test by RT-qPCR, the upper limit of *n* is 100 [1], although other studies indicate an upper limit of 50 for *n* [3-5]. For illustration of the derivation and application of maximum likelihood estimators of *p* and its variance, I will use $N = 1{,}000{,}000; n = 100$, leading to 10,000 tests ($N_{pooled} = 10{,}000$) of the pooled samples.

The raw data from 10,000 tests of pooled samples is condensed to only two numbers, N_{neg} and N_{pos} for the number of negative and positive test results, respectively, together with their associated p_{neg} and p_{pos} (Fig. 1). The expectation of p_{neg} is $\hat{p}_{neg} = (1-p)^n$, where *p* is the prevalence that we need to estimate, so $\hat{p}_{pos} = 1 - \hat{p}_{neg}$. Therefore, the likelihood function for *p* and its logarithm is

$$L = \hat{p}_{Neg}^{N_{Neg}} \hat{p}_{Pos}^{N_{Pos}} = [(1-p)^n]^{N_{Neg}} [1-(1-p)^n]^{N_{Pos}} \quad (1)$$

$$\ln L = N_{Neg}\, ln(\hat{p}_{Neg}) + N_{Pos} ln\,(\hat{p}_{Pos}) \quad (2)$$

The maximum likelihood criterion states that the best p maximizes ln$L$. We take the derivative of ln$L$ with respect to p, set the derivative to 0 and solve for p. This yields the estimator for p as

$$p = 1 - \left(\frac{N_{Neg}}{N_{Pooled}}\right)^{\frac{1}{n}} \quad (3)$$

The variance of p (s_p^2) is the negative inverse of the second derivative of ln$L$ with respect to p, i.e.,

$$\ln L'' = -\frac{N_{Neg}n}{(1-p)^2} - N_{Pos}\frac{(1-p)^n n^2}{(1-p)^2(1-(1-p)^n)} + \frac{N_{Pos}(1-p)^n n}{(1-p)^2(1-(1-p)^n)} - \frac{N_{Pos}((1-p)^n)^2 n^2}{(1-p)^2(1-(1-p)^n)^2} \quad (4)$$

$$s_p^2 = -\frac{1}{lnL''} \quad (5)$$

If $n = 1$, then $s_p^2 = pq/N$ which is our familiar variance estimate when people are tested individually.

For the fictitious data with $N = 1{,}000{,}000; n = 100; N_{neg} = 9000; N_{pos} = 1000$ (Fig. 1), we have $p = 0.001053; s_p^2 = 1.10877221 \times 10^{-9}$. If we had tested all samples individually, then $s_p^2 = 1.05194139 \times 10^{-9}$ given the same p. Thus, the pooled testing strategy has a cost of a slightly increased variance of p, but a benefit of reducing one million individual tests to 10,000 tests of pooled samples.

Acknowledgement

This research was funded by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC, RGPIN/2018-03878) of Canada.

References

1 Mutesa, L. *et al.* A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature* **589**, 276-280, doi:10.1038/s41586-020-2885-5 (2021).

2 Robert, D. The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics* **14**, 436-440, doi:10.1214/aoms/1177731363 (1943).

3 Shental, N. *et al.* Efficient high-throughput SARS-CoV-2 testing to detect asymptomatic carriers. *Science advances* **6**, eabc5961, doi:10.1126/sciadv.abc5961 (2020).

4 Yelin, I. *et al.* Evaluation of COVID-19 RT-qPCR Test in Multi sample Pools. *Clin. Infect. Dis.* **71**, 2073-2078, doi:10.1093/cid/ciaa531 (2020).

5 Lohse, S. *et al.* Pooling of samples for testing for SARS-CoV-2 in asymptomatic people. *Lancet Infect Dis* **20**, 1231-1232, doi:10.1016/s1473-3099(20)30362-5 (2020).

Pooled sample	Individuals	Test result
1	1-100	negative
2	101-200	negative
3	201-300	positive
…	…	…
10000	999901-1000000	negative

N_{neg}	9000
N_{pos}	1000
Sum	10000

p_{neg}	0.9
p_{pos}	0.1
Sum	1

Fig. 1. Illustrative data with 1,000,000 individual samples pooled into 10,000 pooled samples each with 100 individual samples. The 10,000 tests of pooled samples yield 9000 tests positive and 1000 tests negative ($N_{neg} = 9000; N_{pos} = 1000$). $p_{neg} = \frac{N_{neg}}{N}; p_{pos} = \frac{N_{pos}}{N}$