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

License: ☇ ☖ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Missing two key parameters in the pooled testing strategy

Xuhua Xia*

*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 (\(N_{\text{pooled}} = 10,000\)) of the pooled samples.

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

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

\[
\ln L = N_{\text{Neg}}\ln(\hat{p}_{\text{Neg}}) + N_{\text{Pos}}\ln(\hat{p}_{\text{Pos}})
\]
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

\[
p = 1 - \left( \frac{N_{\text{Neg}}}{N_{\text{pooled}}} \right)^{\frac{1}{n}}
\]

(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.,

\[
\ln L'' = -\frac{N_{\text{Neg}}}{(1-p)^2} - N_{\text{pos}} \frac{(1-p)^2 n^2}{(1-p)^2(1-(1-p)^n)^2} + \frac{N_{\text{pos}}(1-p)^n n}{(1-p)^2(1-(1-p)^n)^2} \frac{N_{\text{pos}}((1-p)^n)^2 n^2}{(1-p)^2(1-(1-p)^n)^2} 
\]

(4)

\[
s_p^2 = -\frac{1}{\ln L''}
\]

(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 \( N = 1,000,000; n = 100; N_{\text{neg}} = 9000; N_{\text{pos}} = 1000 \) (Fig. 1), we have \( p = 0.001053; s_p^2 = 1.108777221 \times 10^{-9} \). 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


<table>
<thead>
<tr>
<th>Pooled sample</th>
<th>Individuals</th>
<th>Test result</th>
<th>N_{neg}</th>
<th>N_{pos}</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-100</td>
<td>negative</td>
<td>9000</td>
<td>1000</td>
<td>10000</td>
</tr>
<tr>
<td>2</td>
<td>101-200</td>
<td>negative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>201-300</td>
<td>positive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10000</td>
<td>999901-1000000</td>
<td>negative</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 ($N_{neg} = 9000; N_{pos} = 1000$). $p_{neg} = \frac{N_{neg}}{N}$; $p_{pos} = \frac{N_{pos}}{N}$.