# TN2004-2: Testing for Population Fraction

George Rebane
4 April 2020

We want to determine the fraction of a target population that shares a common attribute, for the detection of which has been developed an unreliable test with known sensitivity and specificity (q.v.). In the sequel we present the methodology for estimating such a fraction within the context of the Covid-19 (C19) pandemic during which various governmental agencies want to know the current fraction of, say, infected people in a target population for the purpose of developing and/or evaluating response policies.

The naïve approach to calculating such a fraction $f_I$ of the infected is to draw a random sample from the population, administer the test to the sample, count the number of 'test positive' (TP) results, and divide by the size of the sample. Then this value of $f_I$ is inferred to the entire population within the standard error bounds of measuring such a fraction from a sample. It turns out that this approach yields estimates with large errors, especially in the range of low $f_I$ values which are most critical for making timely decisions as the disease is spreading through the population. In the sequel we develop the correct method to obtain the desired estimate of the population fraction.

The test in question is determined to have a sensitivity $P(TP|V)$, the probability of yielding a test positive given that the tested individual is infected with the virus V. Its specificity $P(\neg TP|\neg V)$ is the probability that the test returns a test negative given that the individual is not infected. Let $N$ be the size of the random sample drawn from the target population, and $N_{TP}$ be the number of the sample testing positive. The naïve approach yields the estimated infected fraction as

$$f_{TP} = \frac{N_{TP}}{N} \tag{1}$$

But how did that $N_{TP}$ come about? From the test's parameters we can express its expected value as

$$N_{TP} = N_V P\left(TP\,|\,V\right) + \left(N - N_V\right) P\left(TP\,|\,\neg V\right) \ . \tag{2}$$

Unfortunately the actual value of $N_V$ , the actual number of infected in the sample, is unknown. In fact, that is what we would like to determine so as to be able to calculate $f_I$ correctly. We proceed by letting $f_{IA}$ be the actual unknown fraction of the infected in the sample. The expected value of $f_{TP}$ is then given by

$$\begin{aligned} E\left(f_{IA}\right) &= \frac{N_V}{N} P\left(TP\,|\,V\right) + \frac{N - N_V}{N} P\left(TP\,|\,\neg V\right) \\ &= f_{IA} P\left(TP\,|\,V\right) + \left(1 - f_{IA}\right) P\left(TP\,|\,\neg V\right) \simeq f_{TP} \end{aligned} \tag{3}$$

Solving for $f_{IA}$ yields

$$f_{IA} = \frac{f_{TP} - P(TP|\neg V)}{P(TP|V) - P(TP|\neg V)} \ . \tag{4}$$

This is the correct estimate of the actual fraction of the infected people in the sample. At this point it would be instructive to illustrate the difference between $f_{TP}$ and $f_{IA}$. Suppose the actual population infected fraction is $f_{IA} = 0.05$, and we drew a random sample that was tested with a test with performance attributes $P(TP|V) = 0.95$ and $P(TP|\neg V) = 0.10$. Just counting the tests positive and using (1) gives $f_{TP} = 0.1425$ as the 'naïve' infected fraction. Such an approach gives a grossly erroneous answer with an astonishing 185% error. Using (4) would correctly have recovered the desired infected fraction. It is clear from (4) that the naïve fraction equals the correct fraction were the administered test perfectly reliable, that is if $P(TP|V) = 1$ and $P(TP|\neg V) = 0$.

But our analysis does not end here. In addition to using the correct formula to compute the sample's infected fraction, we are also interested in the reliability of that estimate. This involves computing its variance as a function of the errors in its constituents. To begin, for any reasonable $N$ we know that the number of infected $N_V$ it contains is another random variable derived from the $N$ Bernoulli trials in the random selection process from a population whose actual infected fraction is $f_{IA}$ which equals the probability $P_I$ that a randomly drawn person is infected. Constructing such a random sample then dictates that our $N_V$ is defined only to within a binomial probability distribution with expected value $P_I N$ and variance $P_I(1- P_I)N$. From the study of binomial distributions, we also know that for $P_I(1- P_I)N \geq 10$ we are justified in using the gaussian normal distribution instead of the more complex binomial distribution in our analysis. For a sample of $N = 1{,}000$ or more, we have in the above case $P_I(1- P_I)N = 47.5$.

In going forward, we also assume that the sensitivity and specificity of the administered test are known and known to within certain error bounds. This says that we know the test's $P(TP|V)$ and $P(TP|\neg V)$, along with their respective variances. With these data in hand we can now compute the variance of $f_{IA}$ in (4). Before diving into the math, we will relabel the required probabilities for ease of following the development – let $P^+ = P(TP|V)$ and $P^- = P(TP|\neg V)$. Then (4) becomes

$$f_{IA} = \frac{f_{TP} - P^-}{P^+ - P^-} \ , \tag{5}$$

and from the study of error propagation we can write

$$\sigma^2_{f_{IA}} = \left(\frac{\partial f_{IA}}{\partial f_{TP}}\right)^2 \sigma^2_{f_{TP}} + \left(\frac{\partial f_{IA}}{\partial P^+}\right)^2 \sigma^2_{P^+} + \left(\frac{\partial f_{IA}}{\partial P^-}\right)^2 \sigma^2_{P^-} \tag{6}$$

where

$$\frac{\partial f_{IA}}{\partial f_{TP}} = \frac{1}{P^+ - P^-}, \quad \frac{\partial f_{IA}}{\partial P^+} = \frac{P^- - f_{TP}}{\left(P^+ - P^-\right)^2}, \quad \frac{\partial f_{IA}}{\partial P^-} = \frac{f_{TP} - P^+}{\left(P^+ - P^-\right)^2}. \tag{7}$$

For the remainder of our development we assume that the test's performance parameters are normally distributed such that about 95% of the time each probability will lie within ± 0.01 of the mean. This says that we can use $\sigma_{P^+} = \sigma_{P^-} = 0.005$. Also, we can determine from the error propagation of drawing a random sample of size *N* from a population with an $f_{IA}$ that satisfy the above described conditions for using the normal distribution approximation for the underlying binomial distribution. In that case

$$\begin{aligned}
&N_V = f_{IA} N, \quad f_{TP} = N_V \big/ N = f_{IA} \\
&\sigma^2_{N_V} = f_{IA}\left(1 - f_{IA}\right) N, \quad \sigma^2_{f_{TP}} = \left(\frac{\partial f_{TP}}{\partial N_V}\right)^2 \sigma^2_{N_V} \\
&\sigma^2_{f_{TP}} = \left(\frac{1}{N}\right)^2 f_{IA}\left(1 - f_{IA}\right) N = \frac{f_{IA}\left(1 - f_{IA}\right)}{N} \\
&\quad = \frac{(0.05)(0.95)}{1{,}000} = 0.000048 \rightarrow \sigma_{f_{TP}} = 0.0069
\end{aligned} \tag{8}$$

as discussed above. Evaluating (6) with the above inputs yields $\sigma_{f_{IA}} = 0.015$, a marginally tolerable error for the correct estimate of $f_{IA} = 0.05$ from (4). The approximate 95% error bounds for this estimate are ±0.03.

The above development was confirmed through a Monte Carlo simulation with the above parameters. A test run consisted of randomly sampling a population with $f_{IA}$ infected. The random sample, of course, contained a random number as determined by its Bernoulli trial. That sample was tested with a test whose actual sensitivity and specificity were also draws from its performance distributions. That yielded the naïve fraction $f_{TP}$ of test positives, which was then used in (4) to compute $f_{IA}$ which, of course, is also a random number now drawn from a distribution whose theoretical mean and variance are computed from (5) and (6) with the nominal test parameter values known to the tester.

This simulated test was run 10,000 times to generate the histogram of results shown below. The means are shown in red and the plus/minus one sigmas in green (solid from the Monte Carlo runs and dashed for the theoretical values given above. As seen, the means are essentially equal, and the actual sigma is about 0.0017 smaller than the theoretical sigma. This demonstrates that we can expect the experienced distribution to be a bit tighter than that predicted by theory. The theoretical predicted gaussian pdf is shown in black. The bottom line is that the formula in (4) yields an unbiased and correct estimate of the infected fraction of the target population.