mathematics

MDPI

Article

Estimation of Population Prevalence of COVID-19 Using Imperfect Tests

Leonid Hanin

Department of Mathematics and Statistics, Idaho State University, 921 S. 8th Avenue, Stop 8085, Pocatello, ID 83209-8085, USA; hanin@isu.edu; Tel.: +1-208-282-3293

Received: 14 September 2020; Accepted: 27 October 2020; Published: 31 October 2020

Abstract: I formulate three basic biomedical/statistical assumptions that should ideally guide well-designed population prevalence studies of the present or past disease including COVID-19. On the basis of these assumptions alone, I compute several probability distributions required for statistical analysis of testing data collected from a sample of individuals drawn from a heterogeneous population. I also construct a consistent asymptotically unbiased estimator of the population prevalence of the disease or infection from the collected data and derive a simple upper bound for its variance. All the results are rigorously proved and valid for any test for COVID-19 or other disease provided that the sum of the test's sensitivity and specificity is larger than 1. A few recommendations for the design of COVID-19 prevalence studies informed by the results of this work are formulated. The methodology developed in this article may prove applicable to diseases and conditions other than COVID-19 as well as in some non-epidemiological settings.

Keywords: COVID-19; disease prevalence; false positive test result; false negative test result; sensitivity; specificity; study design

MSC: 62P10; 92C60

1. Introduction

Uncovering the true scale of COVID-19 pandemic is critical for shaping public health policy, designing effective medical interventions, and planning economic, social, and educational activities. This requires conducting testing studies aimed at the assessment of the unknown prevalence of the ongoing or past infection in a given population. Knowledge of the prevalence parameter is also indispensable for estimation of the all-important infection mortality rate of COVID-19.

The prevalence of a current or past disease or infection in a population at a given time is defined as the fraction of the population that at the time of interest has or has had the disease or infection. The prevalence is time-dependent. Due to the fact that COVID-19 is highly contagious and has a relatively short incubation period, the prevalence of COVID-19 in a population may change on the time scale of weeks. Thus, to provide a meaningful estimate of the population prevalence of COVID-19, testing studies should be conducted during a short period of time.

Selection of an appropriate target population for a given prevalence study is essential for the study's scientific value. Such selection may be guided by the current knowledge of the epidemiology of the disease. Testing the entire population for COVID-19 or other diseases is expensive and usually impractical. This is why it is typically performed on a relatively small sample of individuals deemed to be representative of the population in terms of the likelihood of a present or past infection. The sample can be drawn randomly or recruited by other means and controlled for individual characteristics known to be associated with the rate of occurrence of the infection or severity of the disease in order to approximate the distribution of these characteristics in the target population. In the case of

COVID-19, these include age, sex, race, socio-economic status, membership in certain professional groups, geographic location, and the presence of various comorbidities. One notable exception to the small sample approach is a SARS-CoV-2 infection study conducted on 21–29 February 2020 and 7 March 2020 in Vo', a municipality of 3275 inhabitants in the Veneto region in Italy, where nasopharyngeal swabs were collected from 85.9% and 71.5% of the entire population in two phases of the study [1].

Several reports on COVID-19 population prevalence studies have been recently published [2–5]. One of them, the Santa Clara county seroprevalence study [2], has stirred a considerable scientific [6] and public [7,8] controversy. The reasons include, but are not limited to, questions raised about the study design, statistical methodology, and the lack of a full disclosure of methods and underlying assumptions. This creates an urgent need for methodological clarity regarding mathematical, statistical, and epidemiological foundations of testing for COVID-19, the analysis of the resulting data including accurate accounting for false positive and false negative test results, and effective design of population prevalence studies. The present work is a step in this direction. In particular, I explicitly state three fairly minimal biomedical/statistical assumptions essential for estimation of the prevalence of COVID-19 or other diseases in a given population. In addition, I identify possible reasons that may lead to violation of these assumptions using the studies [2–5] as an illustration. Finally, detailed proofs of all the mathematical and statistical results are provided. These proofs depend on the three aforementioned assumptions alone, and it is specified which assumptions are requisite for each result.

An important consideration to reckon with when studying prevalence of a present or past infection in a target population is heterogeneity of the infection prevalence across its various subpopulations. These subpopulations may be determined by the above-mentioned *observable* individual characteristics or their combinations. The effects of these attributes on the risk of COVID-19 infection and severity of the disease have been well-documented. For example, health care workers may be exposed to higher doses of SARS-CoV-2, and, more frequently, than members of the general population; consequently, they may have a higher prevalence of the infection. Furthermore, potentially significant fractions of the population may be insusceptible to COVID-19 or carry the infection asymptomatically. In addition, various presently unknown variables associated e.g., with the genetic make-up, functioning of the immune system, and the history of immunizations and exposures to similar pathogens may all prove important for proper prevalence stratification and estimation.

Various approaches to accounting for population heterogeneity can be illustrated using the aforementioned prevalence studies [2–5]. The Gangelt study [3] did not deal with explicitly defined subpopulations; however, it addressed the effects of the household size, the presence of several co-morbidities, and participation in a COVID-19 super-spreading event on the risk of COVID-19 infection and severity of the disease. The Santa Clara county study [2] defined its subpopulations by sex, race, and zip code. To mitigate a mismatch between the sample of tested subjects and the population (manifesting, for example, in the fact that 63.7% of study participants with valid test results were females while the overall proportion of females in the Santa Clara county is 49.5%), the test results were re-weighted, see [2] for details. A similar approach was taken in the Los Angeles county study [4] that stratified the population by sex, age, race/ehtnicity, and income. Finally, the population queried for the prevalence of ongoing COVID-19 infection in the Iceland study [5] was only stratified by age and sex.

A subpopulation of a given population will be called *homogeneous* if, at the time of survey (or over a certain period of time in the case of survey for a past infection), all members of the subpopulation have, or had, the same probability of being infected. This probability represents the prevalence of the disease within the subpopulation. Homogeneous subpopulations of the given population together with their prevalences will be referred to as the *prevalence structure* of the population. Like many other abstract concepts of mathematics and statistics, the concepts of homeogeneous population and prevalence structure are idealizations; however, they will prove to be useful for our analysis.

The diagnostic quality of a test for a certain infection (or disease) is determined by its *sensitivity* α (i.e., the probability of obtaining a positive test result for an infected individual) and *specificity* β (i.e., the probability of a negative test result for an uninfected individual). Great effort is expended by the manufacturers of test kits to ensure that the sensitivity and specificity of their test remain stable across a wide range of testing conditions, which involves running the test on a large number of known true positive and true negative samples. This is why in this article I assume α and β to be fixed. In spite of the effort, however, the sensitivity and specificity of a test, especially a newly designed one, may display certain random or systematic variation depending on the testing site, tested individual, and other conditions. A Bayesian framework for studying the effects of such variation on population prevalence of COVID-19 was presented, in the context of critical analysis of the Santa Clara county study [2], in [6]. For a review of Bayesian modeling approaches to the assessment of accuracy of diagnostic tests, see [9].

Below, the reader will encounter references to various tests for COVID-19 infection including molecular amplification tests such as RT-PCR for SARS-CoV-2 and serological tests for its antibodies, such as IgA, IgM, and IgG that serve as biomarkers of the present or past infection. For a general reference on this subject, see [10].

In this work, I compute in closed form, starting with a homogeneous population, (1) the distribution of the number, n, of positive test outcomes that result from testing N individuals using a test with given sensitivity α and specificity β; (2) the distribution of the number of true and false positive test results conditional on the event that n positive test results were observed; and (3) the conditional expected value and variance of the true number of infected individuals among those tested given n positive test results, see Sections 3 and 4. In Section 5, an *Equivalence Principle* that enables an extension of all these results to heterogeneous populations is established. In Section 6, I bring the above results to bear on the design of a consistent estimator of the population prevalence of the disease (or infection) from the test results collected from a sample of individuals. Finally, in Section 6, I also provide a detailed justification of the mathematical form of the prevalence estimator and study its properties. This estimator was employed in [2] and used for the analysis of uncertainty propagation in [6].

The results of this work can be used for the assessment of plausibility of prevalence estimates reported by various population studies. Specifically, the prevalence of the current or past infection can be checked for self-consistency by comparing the observed number of positive cases with its theoretical prediction and by computing the expected number of false positive and false negative test results and associated probabilities. Additionally, the assumptions formulated in Section 2 may help optimize the design of future prevalence studies for COVID-19 and other diseases.

The Santa Clara county study [2] was criticized for using suboptimal tests for the presence of IgM and IgG antibodies that serve as respective biomarkers of current and past SARS-CoV-2 infection, with combined sensitivity $\alpha = 0.828$ and specificity $\beta = 0.995$ on the grounds that these tests could have produced a large number of false positive and false negative results. I argue that this criticism is largely unfounded and show that the estimator of the true population prevalence, p, derived in the present work and also utilized in [2], is consistent and has small variance for a sufficiently large sample size for *any* α and β such that $\alpha + \beta > 1$ and $1 - \beta < p < \alpha$. For example, if at the time of the study [2], the true COVID-19 seroprevalence in Santa Clara county was, say, between 1% and 10%, then the tests employed (and even those with the same specificity and a dismal sensitivity of 15%) would still be appropriate for a sufficiently large number of tested individuals, should the three assumptions formulated in Section 2 be met.

Because of the current acute interest in SARS-CoV-2 and COVID-19, I will be employing below the terminology and biological considerations pertaining to COVID-19 testing. However, the results of this work may prove applicable to testing for other diseases or conditions as well as in some non-epidemiological settings.

This article can be viewed as a tutorial on some basic mathematical and statistical aspects of testing and prevalence estimation. It represents an expanded and refined version of the author's preprint [11].

2. Basic Assumptions

I start by spelling out three assumptions, termed A, B , C that are ideally to be met by a well-designed population prevalence study. Their discussion is tailored to testing for COVID-19 and uses studies [2–5] as an illustration. As many general assumptions in probability and statistics, assumptions A–C are largely empirically untestable. However, various kinds of empirical evidence as well as careful study design can make the case of their validity stronger. In the terminology of Henri Poincaré [12], these assumptions serve as *neutral hypotheses* that are not too restrictive yet enable building a rigorous quantitative framework for population prevalence estimation. Conversely, specific aspects of study design, discussed below, likely contravene these assumptions. Finally, the assumptions do not have to be necessarily met by the entire sample of subjects recruited for testing or those actually tested; rather, they may serve as a guide for selection of a subset of study participants whose valid test results will be used for population prevalence estimation.

A. Independence. *The events of current or past infection in all tested individuals are stochastically independent.*

Although this assumption is indispensable for rigorous statistical analysis of testing data, its violations in prevalence studies are fairly common. One particular reason is excessive inclusion of multiple members of the same household in a set of testing data used for prevalence estimation. The effects of such oversampling were clearly demonstrated by the study [3] conducted on 31 March–6 April 2020 in Gangelt, a community of around 12,500 people in the state of North Rhine-Westphalia, Germany. One of the aims of the study was to estimate the excess risk of contracting COVID-19 for someone who lives in the same household with an infected individual. It was found that, interestingly, the risk of such secondary infection increases by 28.1%, 20.2%, and 2.8% for households with two, three, and four people, respectively, relative to the 15.5% risk of the primary infection [3]. Given that 919 participants of the Gangelt study for whom valid RT-PCR SARS-CoV-2 test results and/or the titers of IgA or IgG antibodies were obtained belonged to just 405 households, the validity of Assumption A for this study is questionable. The same likely applies to the Santa Clara study [2] conducted on 3–4 April 2020 in Santa Clara county, CA with the population of around 2 million people. In that study, among 3390 participants whose blood specimens were analyzed for the presence of IgM or IgG antibodies there were 2747 adults and 643 children, no more than one per household, living with some of these adults [2]. By contrast to the Gangelt and Santa Clara studies, the Los Angeles study of IgM/IgG seroprevalence [4] that was conducted on 10–11 April 2020 in the Los Angeles county, CA limited participation to one individual per household.

The validity of Assumption A may also prove problematic if a disproportionately large amount of study participants attended a known COVID-19 super-spreading event or may be suspected of belonging to known clusters of COVID-19 cases. For example, investigation of the effects of one such event, a carnival festivity held around 15 February 2020 in Gangelt, revealed that the infection rate among participants of the event was 2.6 higher than among non-participants; in addition, the course of the disease was much more severe in the former than in the latter [3].

B. Test Uniformity. *The sensitivity and specificity of a test are identical for all tested individuals.*

Although this assumption is tacitly adopted almost universally, its validity may prove in certain cases questionable. For example, in asymptomatic or pre-symptomatic carriers of COVID-19, the number of viral particles on a nasopharyngeal swab may be too low to be detectable by the RT-PCR test. Additionally, in some asymptomatic or even symptomatic individuals, the titer of IgA or IgM antibodies indicative of an ongoing disease may not exceed the detection threshold of a serological test. As yet another example, in convalescent COVID-19 patients, the presence of viral particles may

already be undetectable while the titer of IgG antibody, an indicator of a past disease, may not be detectable yet. In all these cases, the sensitivity of the test will be reduced. Likewise, cross-reactivity with viral fragments of, or antibodies to, another virus may increase the likelihood of false positive responses in those individuals who at the time of testing are, or have recently been, infected with a similar pathogen, e.g., a coronavirus causing the common cold. Such cross-reactivity will result in a reduction in the test's specificity.

Another source of systematic non-uniformity of test performance is the use of composite tests. For example, IgM and IgG antibody titers in the Santa Clara county study [2] were measured concurrently, and so were IgA and IgG antibody titers in the Gangelt study. Suppose two tests with respective sensitivities α_1, α_2 and specificities β_1, β_2 were given to the same group of subjects. Whereas the specificity of the composite test for any uninfected individual can be assumed to be $\beta_1\beta_2$,, the sensitivity of the composite test varies. Consider, for example, an infected individual who only has the first antibody. If a positive result is defined as detection of at least one antibody, then, under the independence assumption, the sensitivity of the composite test for such individual would be $\alpha_1 + (1 - \beta_2) - \alpha_1(1 - \beta_2) = 1 - \beta_2(1 - \alpha_1)$. Similarly, for an infected individual who has only the second antibody, it is $1 - \beta_1(1 - \alpha_2)$. However, for those subjects who have both antibodies, the test's sensitivity is $\alpha_1\alpha_2$ (for an evidence that the fraction of such subjects is not negligible, see, e.g., [13]). Thus, the sensitivities of the composite test for these three categories of individuals are, generally speaking, all distinct.

Finally, test results of the Vo' study [1] have not been adjusted for the sensitivity and specificity of the RT-PCR test, which amounts to assuming that $\alpha = \beta = 1$.

C. The Matching Principle. *Tested individuals are selected independently of each other, and the prevalence structure of the sample of these individuals matches that of the target population.*

This assumption suggests a particular way in which the sample of tested individuals is representative of the target population. One sampling method that satisfies the Matching Principle is Simple Random Sampling (SRS), see, e.g., [14], whose defining property is that all samples of a given size are equally likely to be selected. To see that SRS satisfies Assumption C, consider drawing a sample of a given size N from a population of $\mathcal{P}$ individuals. For a fixed subpopulation S, homogeneous or otherwise, consisting of $\mathcal{Q}$ individuals, denote by η the random number of individuals from a sample that belong to S. It follows from the definition of SRS that random variable η has hypergeometric distribution $H(\mathcal{P}, \mathcal{Q}, N)$. Therefore, for its expected value, μ, we have $\mu = N\mathcal{Q}/\mathcal{P}$, so that $\mu/N = \mathcal{Q}/\mathcal{P} = w$, where w is the fractional size, or weight, of the subpopulation S. Thus, under SRS, every subpopulation is represented in the sample, on average, in accordance with its weight.

SRS can be combined with stratification of the population into several subpopulations determined by observable individual characteristics associated with the likelihood of the disease or infection. For example, the total sample size N can be first partitioned into r subsample sizes, $N = N_1 + N_2 + \ldots + N_r$, proportional to the demographic weights of the identified subpopulations and then random subsample of size N_i can be generated from the $i-$th subpopulation by means of SRS for each $i = 1, 2, \ldots, r$.

One source of potential violation of Assumption C is oversampling from the same household, discussed above, allowed by the design of the studies [2,3]. The validity of Assumption C may also prove uncertain if recruitment for testing involves a significant opportunity for self-selection, which makes it likely that people who surmised that they have, or have had, the disease volunteered for the study. Such selection bias was manifestly present in the SARS-CoV-2 population prevalence study [5] conducted in Iceland between 13 March and 1 April 2020 where about half of the 10,797 tested participants who volunteered for the study had mild respiratory symptoms. A possibility for self-selection also existed in the Santa Clara county study whose recruited volunteers responded to an advertisement posted on Facebook [2]. The same problem was potentially present in the Los

Angeles study where only 865 among 1952 randomly selected adults (with some restrictions aimed at matching the county demographics) were actually tested [4]. Finally, the initial recruitment effort of the Gangelt study consisted of generating a random sample of 600 community members with distinct last names and inviting them to participate in the study. However, the 407 study participants who responded to the invitation were allowed to bring in other household members for testing. As a result, 1007 individuals from 405 households were tested [3].

The overall logic and flow of exposition in the rest of the article are as follows. All mathematical results, derived under Assumptions A and B for a homogeneous population, where the probability of having a current or past infection can be assumed the same for all individuals, are formulated in Sections 3 and 4. Next, Section 5 introduces, based on Assumption C, the *Equivalence Principle* that enables a natural extension of all the results obtained in Sections 3 and 4 to a heterogeneous population consisting of any number of homogeneous subpopulations. Section 6 is dedicated to construction of a prevalence estimator and studying its properties. Finally, in Section 7, I summarize the findings and formulate recommendations for the design of prevalence studies informed by the analysis of this article.

3. Distribution of the Number of True and False Positive Test Results

Consider a test with a binary outcome (positive/negative) administered to N individuals selected from a homogeneous population with infection prevalence p, $0 < p < 1$. Let α be the sensitivity and β be the specificity of the test, $0 < \alpha, \beta < 1$. Suppose the test resulted in n, $0 \leq n \leq N$, positive outcomes. Denote by X, Y the respective *unobservable* numbers of true positive and false positive test results, and let $Z = X + Y$ be the *observable* total number of positive outcomes. Denote by M the unknown true number of presently or previously infected individuals (depending on the nature of the test) among the N tested individuals. Below, we seek to compute, under Assumptions A and B, the distribution of random variables X, Y, Z and the conditional distributions of X and Y given $Z = n$. The conditional expectation and variance of random variable M given $Z = n$ will be computed in closed form in the next section.

Assumption A implies that the distribution of random variable M is binomial $B(N, p)$:

$$P(M = m) = \binom{N}{m} p^m (1-p)^{N-m}, \quad 0 \leq m \leq N. \tag{1}$$

For this and other basic concepts and results from probability, the reader is referred to [15].

If $M = m$ is fixed, then the testing of each infected individual produces a positive test result, independently of other individuals (Assumption A), with the same probability α, the sensitivity of the test (Assumption B). Then, for the number, X, of true positive test results, we have

$$P(X = x | M = m) = \binom{m}{x} \alpha^x (1-\alpha)^{m-x}, \quad 0 \leq x \leq m. \tag{2}$$

Similarly, it follows from Assumption B that every uninfected individual receives a false positive test result with the same probability $1 - \beta$. Thus, the distribution of the number, Y, of false positives is given by

$$P(Y = y | M = m) = \binom{N-m}{y} (1-\beta)^y \beta^{N-m-y}, \quad 0 \leq y \leq N-m. \tag{3}$$

Importantly, it follows from Assumption A that, for every m, random variables X and Y are conditionally independent given $M = m$.

Due to Assumptions A and B, random variable X is a *thinning* of the binomial random variable M with probability α. In general, thinning of a sequence of random events is their independent marking, or filtration, with the same probability. Accordingly, the random variable that counts the number of

marked events is called a thinning of the random variable counting the occurrence of the original events; for more on thinning, see [16]. By compounding distributions (1) and (2), we find that random variable X has binomial distribution $B(N, \alpha p)$. In fact, for $0 \leq x \leq N$, we have using the formula of total probability, setting $j = m - x$, and finally employing Newton's binomial formula:

$$P(X = x) = \sum_{m=x}^{N} P(X = x|M = m)P(M = m) = \sum_{m=x}^{N} \binom{m}{x} \alpha^x (1-\alpha)^{m-x} \binom{N}{m} p^m (1-p)^{N-m}$$

$$= (\alpha p)^x \sum_{j=0}^{N-x} \binom{x+j}{x} \binom{N}{x+j} [p(1-\alpha)]^j (1-p)^{N-x-j}$$

$$= \frac{N!}{x!(N-x)!} (\alpha p)^x \sum_{j=0}^{N-x} \frac{(N-x)!}{j!(N-x-j)!} [p(1-\alpha)]^j (1-p)^{N-x-j} = \binom{N}{x} (\alpha p)^x (1-\alpha p)^{N-x}.$$

Likewise, Y represents a thinning of the binomial random variable $N - M$ with probability $1 - \beta$ and consequently has distribution $B[N, (1-\beta)(1-p)]$. Similarly, the distribution of the number of false negative test results is $B[N, (1-\alpha)p]$.

To compute the joint distribution of random variables X and Y, notice that, if $X = x$ and $Y = y$, then every admissible value, m, of random variable M satisfies the inequalities $x \leq m \leq N - y$. Using the formula of total probability, invoking Equations (1)–(3), rearranging the factors, making a change of variable $j = m - x$, and finally employing Newton's binomial formula, we obtain for all $x, y \geq 0$ such that $x + y \leq N$:

$$P(X = x, Y = y) = \sum_{m=x}^{N-y} P(X = x|M = m)P(Y = y|M = m)P(M = m)$$

$$= \sum_{m=x}^{N-y} \binom{m}{x} \alpha^x (1-\alpha)^{m-x} \binom{N-m}{y} (1-\beta)^y \beta^{N-m-y} \binom{N}{m} p^m (1-p)^{N-m}$$

$$= \frac{N!}{x!y!} \alpha^x (1-\beta)^y \sum_{m=x}^{N-y} \frac{(1-\alpha)^{m-x} \beta^{N-m-y}}{(m-x)!(N-m-y)!} p^m (1-p)^{N-m}$$

$$= \frac{N!}{x!y!} (\alpha p)^x [(1-\beta)(1-p)]^y \sum_{j=0}^{N-x-y} \frac{[(1-\alpha)p]^j [\beta(1-p)]^{N-x-y-j}}{j!(N-x-y-j)!}$$

$$= \frac{N!}{x!y!(N-x-y)!} (\alpha p)^x [(1-\beta)(1-p)]^y [(1-\alpha)p + \beta(1-p)]^{N-x-y}. \tag{4}$$

Therefore, random vector $(X, Y, N - X - Y)$ has trinomial distribution

$$Mult[N; \alpha p, (1-\beta)(1-p), (1-\alpha)p + (1-p)\beta].$$

Finally, lumping together true and false positive test results and combining their probabilities lead to a conclusion that the distribution of the total number, $Z = X + Y$, of positive test results, is binomial $B(N, \lambda)$:

$$P(Z = n) = \binom{N}{n} \lambda^n (1-\lambda)^{N-n}, \quad 0 \leq n \leq N, \tag{5}$$

where

$$\lambda = \alpha p + (1-\beta)(1-p) \quad \text{and} \quad 1 - \lambda = (1-\alpha)p + \beta(1-p). \tag{6}$$

According to the formula of total probability, λ represents the probability of obtaining a positive test result for a randomly selected individual from the given homogeneous population.

Formulas (4) and (5) produce the following distributions of the number of true and false positive test results conditional on the observed total number of positive outcomes:

$$P(X = x|Z = n) = \frac{P(X = x, Y = n - x)}{P(Z = n)} = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}, \quad 0 \le x \le n, \tag{7}$$

and

$$P(Y = y|Z = n) = P(X = n - y|Z = n) = \left(\begin{array}{c} n \\ y \end{array} \right) (1-\theta)^y \theta^{n-y}, \quad 0 \le y \le n, \tag{8}$$

where

$$\theta = \frac{\alpha p}{\lambda} = \frac{\alpha p}{\alpha p + (1-\beta)(1-p)} \quad \text{and} \quad 1 - \theta = \frac{(1-\beta)(1-p)}{\alpha p + (1-\beta)(1-p)}. \tag{9}$$

Thus, the distribution of the number of true and false positives given the observed number, n, of positive test results is $B(n, \theta)$ and $B(n, 1-\theta)$, respectively.

Distributions (7) and (8) have the following three notable features:

(a) They are independent of the total number, N, of tested individuals;

(b) Parameters θ and $1 - \theta$ specified in (9) represent the *predictive positive and predictive negative values* that can be obtained by applying Bayes theorem to prior probabilities p and $1 - p$, see, e.g., [17];

(c) Distributions (7) and (8) depend on a *single* parameter

$$\frac{\alpha p}{(1-\beta)(1-p)}$$

that combines the basic parameters p, α, β.

The extraordinary simplicity of Formulas (5), (7), and (8) should not becloud the fact that their validity depends critically on Assumptions A and B.

4. Conditional Expected Number of Infected Individuals for a Given Number of Positive Test Results

A natural estimator, $\hat{p}$, of the prevalence of an infection in a population can be defined as the expected fraction of infected individuals among those tested given the observed number, n, of positive test results:

$$\hat{p} = \frac{\mathbb{E}(M|Z = n)}{N}. \tag{10}$$

The main goal of this section is to compute $\mathbb{E}(M|Z = n)$ in the case of a homogeneous population. For the distribution of random variable M conditional on $Z = n$, we have

$$P(M = m|Z = n) = \frac{P(M = m, Z = n)}{P(Z = n)} = \frac{P(M = m)}{P(Z = n)} \sum_x P(X = x, Y = n - x|M = m)$$

$$= \frac{P(M = m)}{P(Z = n)} \sum_x P(X = x|M = m) P(Y = n - x|M = m), \quad 0 \le m \le N, \tag{11}$$

where $P(Z = n)$ is given in (5)–(6) and x satisfies the inequalities $0 \le x \le m$, $x \le n$ and $n - x \le N - m$ or equivalently

$$a(m, n) := \max\{n + m - N, 0\} \le x \le \min\{m, n\} =: b(m, n). \tag{12}$$

Although (1)–(3) and (5) combined with (11) lead to a formula for the conditional probability $P(M = m|Z = n)$, this formula does not seem to be reducible to a simple expression. However, the corresponding conditional expectation and variance can be computed in closed form, as I show below.

Formula (11) suggests that, in order to find the conditional expectation $\mathbb{E}(M|Z=n)$, one has to compute the following quantity:

$$A(n)=\sum_{m=0}^{N} mP(M=m)\sum_{a(m,n)\le x\le b(m,n)} P(X=x|M=m)P(Y=n-x|M=m),$$

where the bounds for variable x are given in (12). Notice that the range of pairs (x,m) has a simpler representation

$$\{(x,m):\ 0\le x\le n,\ x\le m\le x+N-n\}$$

than for pairs (m,x). Therefore, switching the order of summation, changing the variable in the internal sum to $j=m-x$, and using Formulas (1)–(3) yield

$$A(n)=\sum_{x=0}^{n}\sum_{m=x}^{N-n+x} mP(M=m)P(X=x|M=m)P(Y=n-x|M=m)$$

$$=\sum_{x=0}^{n}\sum_{j=0}^{N-n}(x+j)P(M=x+j)P(X=x|M=x+j)P(Y=n-x|M=x+j) \tag{13}$$

$$=\sum_{x=0}^{n}\sum_{j=0}^{N-n}(x+j)\begin{pmatrix} N \\ x+j\end{pmatrix}p^{x+j}(1-p)^{N-x-j}\begin{pmatrix} x+j \\ x\end{pmatrix}\alpha^{x}(1-\alpha)^{j}\begin{pmatrix} N-x-j \\ n-x\end{pmatrix}(1-\beta)^{n-x}\beta^{N-n-j}$$

$$=\frac{N!}{(N-n)!}\sum_{x=0}^{n}\frac{(\alpha p)^{x}[(1-\beta)(1-p)]^{n-x}}{x!(n-x)!}\sum_{j=0}^{N-n}(x+j)\begin{pmatrix} N-n \\ j\end{pmatrix}[(1-\alpha)p]^{j}[\beta(1-p)]^{N-n-j}.$$

Using (6) we represent the internal sum in (13) as

$$(1-\lambda)^{N-n}\sum_{j=0}^{N-n}(x+j)\begin{pmatrix} N-n \\ j\end{pmatrix}\delta^{j}(1-\delta)^{N-n-j}=(1-\lambda)^{N-n}[x+(N-n)\delta],$$

where we set $\delta=(1-\alpha)p/(1-\lambda)$ and used the formula for the expected value of the binomial distribution $B(N-n,\delta)$. Now, the above derivation of the formula for $A(n)$ can be continued:

$$A(n)=\begin{pmatrix} N \\ n\end{pmatrix}(1-\lambda)^{N-n}\sum_{x=0}^{n}\begin{pmatrix} n \\ x\end{pmatrix}(\alpha p)^{x}[(1-\beta)(1-p)]^{n-x}[x+(N-n)\delta]$$

$$=\begin{pmatrix} N \\ n\end{pmatrix}\lambda^{n}(1-\lambda)^{N-n}\sum_{x=0}^{n}\begin{pmatrix} n \\ x\end{pmatrix}\theta^{x}(1-\theta)^{n-x}[x+(N-n)\delta]=P(Z=n)[n\theta+(N-n)\delta],$$

where we employed (5) along with the formula for the expected value of the binomial distribution $B(n,\theta)$, where $\theta=\alpha p/\lambda$ is the same as in Formula (9). Thus, in view of (11),

$$\mathbb{E}(M|Z=n)=\frac{A(n)}{P(Z=n)}=n\theta+(N-n)\delta=n\frac{\alpha p}{\lambda}+(N-n)\frac{(1-\alpha)p}{1-\lambda}. \tag{14}$$

A very similar argument leads to the following formula for the conditional second moment of M given $Z=n$:

$$\mathbb{E}(M^2|Z=n)=P(Z=n)\left\{n\theta[1+(n-1)\theta]+2n(N-n)\theta\delta+(N-n)\delta[1+(N-n-1)\delta]\right\}.$$

Therefore, due to (14) and (6),

$$Var(M|Z=n)=\mathbb{E}(M^2|Z=n)-[\mathbb{E}(M|Z=n)]^2$$

$$= \{n\theta[1+(n-1)\theta] + 2n(N-n)\theta\delta + (N-n)\delta[1+(N-n-1)\delta]\} - [n\theta + (N-n)\delta]^2 \quad (15)$$

$$= n\theta(1-\theta) + (N-n)\delta(1-\delta) = p(1-p)\left[\frac{\alpha(1-\beta)}{\lambda^2}\, n + \frac{(1-\alpha)\beta}{(1-\lambda)^2}\,(N-n)\right].$$

Inspection of Formulas (14) and (15) reveals that the conditional expectation and variance of random variable M given $Z = n$ depend on the following two combinations of parameters p, α, β alone:

$$\frac{\alpha p}{(1-\beta)(1-p)} \quad \text{and} \quad \frac{(1-\alpha)p}{\beta(1-p)}.$$

5. The Equivalence Principle for Heterogeneous Populations

Sections 3 and 4 dealt with a population that was assumed homogeneous in the sense that all its individuals had the same probability, p, to have a current or past infection. The aim of this section is to extend the results of Sections 3 and 4 to a more realistic case of a heterogeneous population consisting of r homogeneous subpopulations. Let $\mathbf{w} = (w_1, w_2, \ldots, w_r)$, where $\sum_{i=1}^r w_i = 1$, is the vector of relative sizes (weights) of these subpopulations and $\mathbf{p} = (p_1, p_2, \ldots, p_r)$ be the vector of their disease prevalences.

I start with introducing the following convenient notation. For a non-negative integer vector $\mathbf{x}$ with r components, set $\mid \mathbf{x} \mid = \sum_{i=1}^r x_i$ and $\mathbf{x}! = \Pi_{i=1}^r x_i!$. In addition, for two such vectors $\mathbf{x}$ and $\mathbf{y}$, we denote $\mathbf{x}\cdot\mathbf{y} = \sum_{i=1}^r x_i y_i$, $\mathbf{xy} = (x_1y_1, x_2y_2, \ldots, x_ry_r)$ and $\mathbf{x}^{\mathbf{y}} = \Pi_{i=1}^r x_i^{y_i}$. Finally, $\mathbf{y} \le \mathbf{x}$ means that $y_i \le x_i$ for $i = 1, 2, \ldots, r$.

Let N_i, $1 \le i \le r$, be the number of individuals from the $i-$th homogeneous subpopulation among N tested individuals. The Matching Principle (Assumption C) implies that random vector $\mathbf{N} = (N_1, N_2, \ldots, N_r)$ with $\mid \mathbf{N} \mid = N$ has multinomial distribution $Mult(N, \mathbf{w})$:

$$P(\mathbf{N} = \mathbf{x}) = \frac{N!}{\mathbf{x}!}\,\mathbf{w}^{\mathbf{x}}, \quad \mid \mathbf{x} \mid = N. \quad (16)$$

Let M_i be the number of infected individuals among N_i tested individuals, $1 \le i \le r$. Random vector $\mathbf{M} = (M_1, M_2, \ldots, M_r)$ represents a component-wise thinning of random vector $\mathbf{N}$ with thinning probabilities forming the vector $\mathbf{p}$. What is the distribution of random vector $\mathbf{M}$? A computation below shows that, in contrast to the binomial case, it is not multinomial!

It follows from Assumption A and subpopulation homogeneity that, for any i, $1 \le i \le r$,, the conditional distribution of random variable M_i given $N_i = x_i$ is binomial $B(x_i, p_i)$. In addition, Assumption A implies that, for every vector $\mathbf{x}$ such that $\mid \mathbf{x} \mid = N$, components of random vector $\mathbf{M}$ are conditionally independent given $\mathbf{N} = \mathbf{x}$. Therefore, for any vector $\mathbf{y}$ with $\mid \mathbf{y} \mid \le N$, we have using (16), employing the formula of total probability, making a change of variable $\mathbf{z} = \mathbf{x} - \mathbf{y}$, and, finally using the multinomial formula,

$$P(\mathbf{M} = \mathbf{y}) = \sum_{\mathbf{x}\ge\mathbf{y},\ |\mathbf{x}|\ =N} P(\mathbf{M} = \mathbf{y}|\mathbf{N} = \mathbf{x})P(\mathbf{N} = \mathbf{x})$$

$$= \sum_{\mathbf{x}\ge\mathbf{y},\ |\mathbf{x}|\ =N} \frac{\mathbf{x}!}{\mathbf{y}!(\mathbf{x}-\mathbf{y})!}\,\mathbf{p}^{\mathbf{y}}(\mathbf{1}-\mathbf{p})^{\mathbf{x}-\mathbf{y}}\frac{N!}{\mathbf{x}!}\,\mathbf{w}^{\mathbf{x}} = \frac{N!}{\mathbf{y}!}(\mathbf{wp})^{\mathbf{y}} \sum_{|\mathbf{z}|\ =N-|\mathbf{y}|} \frac{[\mathbf{w}(\mathbf{1}-\mathbf{p})]^{\mathbf{z}}}{\mathbf{z}!} \quad (17)$$

$$= \frac{N!}{(N-\mid \mathbf{y} \mid)!}\,\frac{(\mathbf{wp})^{\mathbf{y}}}{\mathbf{y}!}\left[\sum_{i=1}^r w_i(1-p_i)\right]^{N-|\mathbf{y}|} = \frac{N!}{(N-\mid \mathbf{y} \mid)!}(1-\mathbf{w}\cdot\mathbf{p})^{N-|\mathbf{y}|}\,\frac{(\mathbf{wp})^{\mathbf{y}}}{\mathbf{y}!}.$$

Due to Assumption B, all the computations in Sections 3 and 4 involve only the *total* number, $m = \mid \mathbf{M} \mid$, of infected individuals among those tested. The distribution of random variable $\mid \mathbf{M} \mid$ can now be derived using the multinomial formula and (17):

$$P(\mid \mathbf{M} \mid = m) = \sum_{|\mathbf{y}|=m} \frac{N!}{(N- \mid \mathbf{y} \mid)!}(1 - \mathbf{w} \cdot \mathbf{p})^{N-|\mathbf{y}|} \frac{(\mathbf{wp})^{\mathbf{y}}}{\mathbf{y}!}$$

$$= \frac{N!}{(N-m)!}(1 - \mathbf{w} \cdot \mathbf{p})^{N-m} \sum_{|\mathbf{y}|=m} \frac{(\mathbf{wp})^{\mathbf{y}}}{\mathbf{y}!} = \frac{N!}{m!(N-m)!}(1 - \mathbf{w} \cdot \mathbf{p})^{N-m} \left(\sum_{i=1}^{r} w_i p_i \right)^m \tag{18}$$

$$= \begin{pmatrix} N \\ m \end{pmatrix} (1 - \mathbf{w} \cdot \mathbf{p})^{N-m} (\mathbf{w} \cdot \mathbf{p})^m, \quad 0 \leq m \leq N.$$

Thus, the total number of infected individuals among N subjects tested follows the binomial distribution $B(N, \mathbf{w} \cdot \mathbf{p})$.

Comparison between Formulas (18) and (1) leads to the following conclusion that can be termed the **Equivalence Principle:**

Under Assumption C, the distribution of the total number of infected individuals among N tested individuals selected from a heterogeneous population consisting of r homogeneous subpopulations with weights $w_1, w_2, \ldots, w_r$ *and infection prevalences* $p_1, p_2, \ldots, p_r$ *is the same as for a homogeneous population with infection prevalence*

$$p = \sum_{i=1}^{r} w_i p_i. \tag{19}$$

The Equivalence Principle is also true when the r subpopulations comprising the population of interest are heterogeneous. In fact, partitioning them into homogeneous subsubpopulations, applying Formula (19), regrouping and rescaling the terms pertaining to the same subpopulation, and applying the Equivalence Principle again leads to Formula (19) in which w_i are the weights (relative sizes) and p_i are the prevalences of the r heterogeneous subpopulations.

In summary, all the results in Sections 3 and 4 that were derived for a homogeneous population with infection prevalence p are also valid for a heterogeneous population if one selects p in accordance with Formula (19). This equivalence property is, of course, quite natural; however, it depends on Assumptions A–C in very essential ways.

6. Prevalence Estimation

If N individuals drawn from a population of interest were tested and n positive test results were observed, then a "naïve" estimate of the prevalence of the current or past infection in the population would be $p_0 = n/N$. The testing process can be viewed as the following mental experiment: for an infected individual, a coin is flipped that lands "heads" with probability α and "tails" with probability $1 - \alpha$ while, for an uninfected individual, another coin is flipped that lands "heads" with probability $1 - \beta$ and "tails" with probability β. In these terms, p_0 is the fraction of "heads" (positive test results) recorded for N independent replications of this random experiment. Clearly, p_0 depends on the sensitivity and specificity of the test and the prevalence of the disease. Therefore, one needs to untangle them and construct a consistent estimator of the prevalence alone.

In the rest of this section, it will be assumed that

$$\alpha + \beta > 1. \tag{20}$$

This condition is always met in practice; otherwise, either the sensitivity or the specificity of a test would not exceed 0.5, thus making the test equivalent or inferior, for either infected or uninfected individuals, to flipping a fair coin.

Recall that the number, n, of positive test results has binomial distribution $B(N, \lambda)$, see Section 3, where $\lambda = \alpha p + (1-\beta)(1-p)$. Notice that, under condition (20),

$$1 - \beta < \lambda < \alpha. \tag{21}$$

I first define the desired estimate, $\hat{p}$, of the population prevalence p heuristically. Because p_0 is a consistent unbiased estimator of λ, the following "plug-in" equation can be set up for $\hat{p}$:

$$\alpha\hat{p} + (1-\beta)(1-\hat{p}) = p_0. \tag{22}$$

This defines $\hat{p}$ as the population prevalence that would produce, on average, the same fraction of positive test results when N individuals are tested as the one actually observed. Solving Equation (22) for $\hat{p}$ yields

$$\hat{p} = \frac{p_0 + \beta - 1}{\alpha + \beta - 1}. \tag{23}$$

Note that this estimator was employed in the Santa Clara county study [2], see also [6]. Observe that $0 < \hat{p} < 1$ if and only if $1 - \beta < p_0 < \alpha$; compare with (21). Thus, the complete definition of estimator $\hat{p}$ is

$$\hat{p} = \begin{cases} 0 & \text{if } \ p_0 \le 1-\beta \\ \frac{p_0+\beta-1}{\alpha+\beta-1} & \text{if } 1-\beta < p_0 < \alpha \\ 1 & \text{if } \ p_0 \ge \alpha \end{cases}$$

This formula implies that meaningful prevalence estimation in a population with low prevalence of a disease requires a test with high specificity (namely, with $\beta > 1 - p_0$, where p_0 is the raw positivity rate for a representative sample). Likewise, estimation of large prevalence requires a test with sufficiently high sensitivity (specifically, with $\alpha > p_0$).

Since $p_0 \to \lambda$ almost surely as $N \to \infty$, Equations (23), (21), and (6) imply that

$$\hat{p} \to \frac{\lambda + \beta - 1}{\alpha + \beta - 1} = p$$

almost surely as $N \to \infty$. Therefore, $\hat{p}$ is a consistent estimator of p. Because estimator $\hat{p}$ is uniformly bounded, we also have $\mathbb{E}\hat{p} \to p$ as $N \to \infty$, which means that estimator $\hat{p}$ is asymptotically unbiased.

The heuristic formula (23) can be derived on more "theoretical" grounds. Recall that $\hat{p}$ is defined as the expected fraction of infected individuals among those tested conditional on the observed number of positive test results, see Equation (10). Then, in view of (14),

$$f(p) := \frac{\mathbb{E}(M|Z=n)}{N} = \frac{\alpha p}{\lambda} p_0 + \frac{(1-\alpha)p}{1-\lambda}(1-p_0).$$

The principal difficulty with this definition of the prevalence estimator $\hat{p}$ is that it depends on the unknown true prevalence parameter p that $\hat{p}$ seeks to estimate. A natural idea, then, would be to determine the value of p for which $f(p) = p$ and take it as the desired prevalence estimator. Using expression (6) for λ, one finds after some algebra

$$f(p) - p = \frac{p(1-p)(1-\alpha-\beta)[p(\alpha+\beta-1)+1-\beta-p_0]}{\lambda(1-\lambda)}.$$

Upon comparison with (23), this leads to the conclusion that, under assumption (20), the required fixed point of function f is exactly the above heuristic estimator $\hat{p}$!

My next goal is to estimate the variance of $\hat{p}$. Recall that a function $T : \mathbb{R} \to \mathbb{R}$ is called a *contraction* if $\mid T(x) - T(y) \mid \le \mid x - y \mid$ for all $x, y \in \mathbb{R}$. Setting here $y = 0$ shows that for every contraction T

$$\mid T(x) \mid \le \mid T(0) \mid \ + \mid x \mid . \tag{24}$$

An important family of contractions consists of functions $T_{a,b}$ defined for $a < b$ by

$$T_{a,b}(x) = \begin{cases} a & \text{if } x \leq a \\ x & \text{if } a < x < b \\ b & \text{if } x \geq b \end{cases}$$

If U is a random variable with finite second moment defined on a sample space S with probability measure P and $V = T(U)$, where T is a contraction, then it follows from (24) that the second moment of random variable V is also finite. Moreover, $VarV \leq VarU$. In fact,

$$VarV = \frac{1}{2}\int\int_{S\times S}[V(s) - V(t)]^2 dP(s)dP(t) = \frac{1}{2}\int\int_{S\times S} \mid T[U(s)] - T[U(t)] \mid^2 dP(s)dP(t)$$

$$\leq \frac{1}{2}\int\int_{S\times S} \mid U(s) - U(t) \mid^2 dP(s)dP(t) = VarU.$$

In particular, setting

$$U = \frac{p_0 + \beta - 1}{\alpha + \beta - 1},$$

we conclude from the above full definition of estimator $\hat{p}$ that $\hat{p} = T_{0,1}(U)$. This leads to the following upper bound for the variance of $\hat{p}$:

$$Var\hat{p} \leq VarU = \frac{Varp_0}{(\alpha + \beta - 1)^2} = \frac{\lambda(1-\lambda)}{N(\alpha + \beta - 1)^2} \leq \frac{1}{4N(\alpha + \beta - 1)^2}. \tag{25}$$

Interestingly, this upper bound depends only on the quantity $\sqrt{N}(\alpha + \beta - 1)$.

The estimator $\hat{p}$ has a remarkable feature that can be termed the "mixture-invariance" property. Consider a population that consists of r subpopulations. Let $N_i \geq 1$ be the number of tested individuals from the $i-$th subpopulation and n_i be the number of positive outcomes, based on the same test. Denote by $\hat{p}_i$ the above-designed prevalence estimate for the $i-$th subpopulation and set $\hat{w}_i = N_i/N$, where $N = N_1 + N_2 + \cdots + N_r$. Finally, assume that $1 - \beta < n_i/N_i < \alpha$ for all i. Then, the prevalence estimate, $\hat{p}$, for the entire population becomes

$$\hat{p} = \sum_{i=1}^{r} \hat{w}_i \hat{p}_i, \tag{26}$$

compared with (19).

In fact, it follows from Formula (23) and $\sum_{i=1}^{r} \hat{w}_i = 1$ that

$$\sum_{i=1}^{r} \hat{w}_i \hat{p}_i = \sum_{i=1}^{r} \hat{w}_i \frac{\frac{n_i}{N_i} + \beta - 1}{\alpha + \beta - 1} = \frac{\sum_{i=1}^{r} \hat{w}_i \frac{n_i}{N_i} + \beta - 1}{\alpha + \beta - 1}. \tag{27}$$

Observe that

$$\sum_{i=1}^{r} \hat{w}_i \frac{n_i}{N_i} = \sum_{i=1}^{r} \frac{N_i}{N} \cdot \frac{n_i}{N_i} = \frac{1}{N}\sum_{i=1}^{r} n_i = \frac{n}{N} = p_0.$$

Thus, p_0 is a weighted average of the ratios n_i/N_i, which implies, due to the above-assumed bounds for these ratios that $1 - \beta < p_0 < \alpha$. We now continue (27) to finally get

$$\sum_{i=1}^{r} \hat{w}_i \hat{p}_i = \frac{p_0 + \beta - 1}{\alpha + \beta - 1} = \hat{p}.$$

The mixture-invariance property (26) enables one to combine prevalence estimates for several subpopulations of a given population obtained within the same study, or different studies utilizing the same test, into a prevalence estimate for the entire population.

7. Discussion and Recommendations

In this article, I computed the distribution of the number of positive outcomes resulting from administration of a test with known sensitivity and specificity to N individuals selected from a given population. I also found the conditional distribution of the unobservable number of true and false positive test results given the observed number, n, of positive outcomes. These formulas lead to a closed form expression for the expected value of the unknown true number of infected individuals among those tested conditional on n. In Sections 3 and 4, these results were obtained for a homogeneous population while the *Equivalence Principle* derived in Section 5 extended them to a heterogeneous population. This theory culminated with a construction in Section 6 of a consistent estimator, $\hat{p}$, of the prevalence of infected individuals in a population and finding an upper bound for the variance of $\hat{p}$.

Importantly, in Section 2, I formulated three basic assumptions required for the validity of the above results and identified, using the well-known early COVID-19 prevalence studies [2–5] as examples, several sources of their violation. Because it is uncommon for epidemiological studies including [2–5] to disclose all the details of their statistical analyses, it is hard to say with certainty if some of the formulas obtained in this work were employed in the published prevalence studies (as mentioned earlier, the Santa Clara county study [2] did use the prevalence estimator (23)). However, the design of these studies does not seem to fully meet the assumptions upon which these formulas vitally depend.

Note that this work never employed any asymptotic arguments, i.e., those applicable to large values of N. Therefore, all the results can be used for small sample size N as long as the sample of tested individuals is sufficiently representative of the prevalence structure of the target population.

One possible line of extension of this work would be to incorporate the compliance rate of individuals recruited for a prevalence study and the rate at which a testing system produces invalid results into the prevalence estimator. Another direction of further work would be to develop a probabilistic framework for the propagation of uncertainty in the test's sensitivity and specificity into the population prevalence estimator. As mentioned in the Introduction, a Bayesian approach to quantifying such propagation was developed in [6].

I close with a list of specific conclusions and recommendations regarding the design of population prevalence studies and relevant statistical methodology.

1. The estimator, $\hat{p}$, of population prevalence p introduced in Section 6 is consistent for *any* test whose sensitivity, α, and specificity, β, satisfy the conditions $\alpha + \beta > 1$ and $1 - \beta < p < \alpha$. The quality of this estimator, as determined by the magnitude of its variance, depends on the quantity $\sqrt{N}(\alpha + \beta - 1)$ alone, see inequality (25), which can be used for deciding on the number of individuals to be tested. While the accuracy of estimator $\hat{p}$ improves with the increase in the sensitivity and specificity of the test, the same improvement can be achieved by increasing the sample size. Thus, contrary to the common belief, high sensitivity and specificity of a test is not of primary importance for population prevalence estimation (although the accuracy of *individual* test results depends critically on how close the test's sensitivity and specificity are to 100%).

2. The "naïve" prevalence estimator $p_0 = n/N$ depends on the true population prevalence and the test's sensitivity and specificity. It may deviate considerably from $\hat{p}$, the correct "disentangled" population prevalence estimator. For example, for a perfectly specific test ($\beta = 1$), one has $\hat{p} = \min\{p_0/\alpha, 1\}$. Therefore, the use of p_0 as prevalence estimator should be discouraged, unless a test with sensitivity and specificity close to 100% is employed.

3. Accurate prevalence estimation for a population with a high prevalence of a disease or infection requires a very sensitive test while, for a population with low prevalence, a very specific test should be used.

4. Prevalence estimates, resulting from data obtained on the same testing platform, for subpopulations with known weights leads automatically, through the "mixing-invariance" property, see Section 6, to a prevalence estimate for a heterogeneous population comprised of these subpopulations without the need for a de novo study.

5. The validity of an estimate of population prevalence of the current or past infection depends critically on study design. Here are a few recommendations for a selection of study participants and choosing among them those individuals whose valid test results can be used for prevalence estimation:

(a) Excessive inclusion of testing data for more than one household member in the same analysis of the prevalence of COVID-19 should be avoided. The same applies to individuals who are known to have been in close and/or protracted contact without PPE.

(b) Prevalence studies in populations where known COVID-19 super-spreading events have occurred should eschew oversampling from the subpopulation of their participants. Likewise, study participants could be screened for association with known infection clusters.

(c) Compliance with the Matching Principle (Assumption C) can be achieved through simple random sampling from the target population. This approach can be combined with stratification based on individual characteristics relevant to the prevalence of the disease or infection. Selected individuals can be additionally screened based on the above criteria (a) and (b) as well as for membership in high-risk professional groups and for residence in locations with high or low prevalence of registered cases.

(d) Self-selection of study subjects can make the resulting prevalence estimate unreliable. One way to prevent this selection bias is to improve post-selection compliance by providing financial incentives to study participants.

(e) Using composite testing data (e.g., combining molecular RT-PCR or IgA/IgM serological test for current infection with serological IgG test for past infection) within the same prevalence analysis violates the testing uniformity assumption and should be avoided.

Author Contributions: L.H. is responsible for all components of the manuscript and all aspects of its preparation. The author has read and agreed to the published version of the manuscript.

Funding: The author confirms that no funding associated with this manuscript was provided.

Acknowledgments: The author thanks Boris Hanin for his help with formatting this article and fruitful discussion of its content.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Lavezzo, E.; Franchin, E.; Ciavarella, C.; Cuomo-Dannenburg, G.; Barzon, L.; Del Vecchio, C.; Rossi, L.; Manganelli, R.; Loregian, A.; Navarin, N.; et al. Suppression of a SARS-CoV-2 outbreak in the Italian municipality of Vo'. *Nature* **2020**, *584*, 425–429. [PubMed]
2. Bendavid, E.; Mulaney, B.; Sood, N.; Shah, S.; Ling, E.; Bromley-Dulfano, R.; Lai, C.; Weissberg, Z.; Saavedra-Walker, R.; Tedrow, J.; et al. COVID-19 antibody seroprevalence in Santa Clara county, California. *MedRxiv* **2020**. [CrossRef]
3. Streeck, H.; Schulte, B.; Kümmerer, B.M.; Richter, E.; Höller, T.; Fuhrmann, C.; Bartok, E.; Dolscheid, R.; Berger, M.; Wessendorf, L.; et al. Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event. *MedRxiv* **2020**. [CrossRef]
4. Sood, N.; Simon, P.; Ebner, P.; Eichner, D.; Reynolds, J.; Bendavid, E.; Bhattacharya, J. Seroprevalence of SARS-CoV-2 – specific antibodies among adults in Los Angeles county, California, on 10–11 April 2020. *J. Am. Med. Assoc.* **2020**, *323*, 2425–2427. [CrossRef]
5. Gudbjartsson, D.F.; Helgason, A.; Jonsson, H.; Magnusson, O.T.; Melsted, P.; Norddahl, G.L.; Saemundsdottir, J.; Sigurdsson, A.; Sulem, P.; Agustsdottir, A.B.; et al. Spread of SARS-CoV-2 in the Icelandic population. *N. Engl. J. Med.* **2020**, *382*, 2302–2315. [CrossRef] [PubMed]
6. Gelman, A; Carpenter, B. Bayesian analysis of tests with unknown specificity and sensitivity. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **2020**. [CrossRef]

7. Krieger, L.M. Feud over Stanford Coronavirus Study: 'The Authors Owe us all an Apology.' Angry Statisticians Dispute Santa Clara County Research. Available online: https://www.mercurynews.com/2020/04/20/feud-over-stanford-coronavirus-study-the-authors-owe-us-all-an-apology/ (accessed on 20 April 2020).
8. Schulson, M. On COVID-19, a Respected Science Watchdog Raises Eyebrows. Available online: https://www.medscape.com/viewarticle/929920 (accessed on 4 May 2020).
9. Branscum, A.J.; Gardner, I.A.; Johnson, W.O. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Prev. Vet. Med.* **2005**, *68*, 145–163. [CrossRef] [PubMed]
10. Turgeon, M. *Immunology and Serology*, 6th ed.; Elsevier: St. Lois, MO, USA, 2017.
11. Hanin, L. The mathematics of testing with application to prevalence of COVID-19. *MedRxiv* **2020**. [CrossRef]
12. Poincaré, H. *Science and Hypothesis*; Dover Publications: New York, NY, USA, 1952.
13. Long, Q.-X.; Liu, B.-Z.; Deng, H.-J.; Wu, G.-C.; Deng, K.; Chen, Y.-K.; Liao, P.; Qiu, J.F.; Lin, Y.; Cai, X.F. Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nat. Med.* **2020**, *26*, 845–848. [CrossRef] [PubMed]
14. Agresti, A. *Categorical Data Analysis*, 3rd ed.; John Wiley and Sons: New York, NY, USA, 1990.
15. Ross, S. *A First, Course in Probability*, 9th ed.; Pearson Education: Boston, MA, USA, 2014.
16. Harremoës, P.; Johnson, O.; Kontoyiannis, I. Thinning and the Law of Small Numbers. In Proceedings of the 2007 IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 1491–1495.
17. Altman, D.G; Bland, J.M. Diagnostic tests 2: Predictive values. *Br. Med. J.* **1994**, *309*, 102. [CrossRef] [PubMed]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).