KIEL INSTITUTE FOR THE WORLD ECONOMY

KIEL WORKING PAPER

Estimating the Prevalence

Properties of the estimator regarding specificity and sensitivity of the underlying test

No. 2152 March 2020

Jens Boysen-Hogrefe, Vincent Stamer

Kiel Institute for the World Economy
ISSN 1862–1155

Leibniz Association

ABSTRACT

ESTIMATING THE PREVALENCE: PROPERTIES OF THE ESTIMATOR REGARDING SPECIFICITY AND SENSITIVITYOF THE UNDERLYING TEST

Jens Boysen-Hogrefe and Vincent Stamer

We provide a calculation tool to assess the properties of a maximum-likelihood (ML) estimator that extrapolates the true prevalence of an infectious disease from a random sample. The tools allow the researcher to correct for the specificity and sensitivity of the underlying medical test, calculate the standard deviation of the estimator and to plan the needed sample size. This document explains the underlying methods of the calculation tools and provides instructions for their proper use. We apply an adaption of the epidemiological SEIR-model to show that ML-estimators from random sampling tests provide a more realistic rate of infection than common approaches.

Keywords: Infectious diseases, random sampling, maximum-likelihood, SEIR-model

JEL classification:

Jens Boysen-Hogrefe
Kiel Institute for the World Economy
Kiellinie 66
D-24105 Kiel, Germany
Email:
jens.hogrefe@ifw-kiel.de
www.ifw-kiel.de

Vincent Stamer
Kiel Institute for the World Economy
Kiellinie 66
D-24105 Kiel, Germany
Email:
vincent.stamer@ifw-kiel.de
www.ifw-kiel.de

Estimating the prevalence: Properties of the estimator regarding specificity and sensitivity of the underlying test*

Jens Boysen-Hogrefe† Vincent Stamer ‡

March 26, 2020

Abstract

We provide a calculation tool to assess the properties of a maximum-likelihood (ML) estimator that extrapolates the true prevalence of an infectious disease from a random sample. The tools allow the researcher to correct for the specificity and sensitivity of the underlying medical test, calculate the standard deviation of the estimator and to plan the needed sample size. This document explains the underlying methods of the calculation tools and provides instructions for their proper use. We apply an adaption of the epidemiological SEIR-model to show that ML-estimators from random sampling tests provide a more realistic rate of infection than common approaches.

*The authors greatly appreciate the research assistance of Falk Wendorff

†Kiel Institute for the World Economy

‡Kiel Institute for the World Economy, Kiel University

1 Aim of the calculators

During pandemic outbreaks of infectious diseases policy makers are forced to take actions against the spread quickly, often at the expense of economic activity. A recent study by Burns et al. (2006) estimates that 60% of economic damages incurred during a pandemic can be attributed to demand shocks, i.e. the indirect costs of an outbreak. Factoring in the interruption of supply chains and detrimental uncertainty likely increases the economic costs significantly. While human health must be protected, governments typically have limited information on the actual spread of the disease. Indeed, the true rate of infection in the population is rarely known. Random testing can be a remedy to achieve the needed information of the prevalence. However, given that specificity and sensitivity of a test can deviate from one, the prevalence has to be estimated from test results e.g. via a Maximum Likelihood estimation.

We provide the ready-to-use tools for such a Maximum Likelihood estimation, which calculates the standard deviation of the estimator for given sensitivity, specificity, sample size and expected prevalence. Vice versa the needed sample size can be retrieved for a standard deviation or precision that shall be achieved.

While the tools presented in this paper are applicable to any infectious disease, we provide examples from the COVID-19 pandemic. Indeed, the outbreak of Sars-CoV-2 is a suitable illustration for the need of statistical tests: At the moment, mainly patients who are at high risk of infection (e.g. because of contact with an infected individual) are tested for the presence of the pathogen by use of a rRT-PCR (reverse transcription polymerase chain reaction) test. This approach swiftly diagnoses COVID-19 and helps to trace the chain of infection. However, the virus has a high level of contagion, an incubation period of approximately five days (Lauer et al., 2020) and results only in minor symptoms for many people. This suggests that the true rate of

infection may be a multiple of the observed rate of infection. This makes it difficult for epidemiologists and policy makers to ascertain "herd immunity" or select the health measures that are most appropriate to balance human health and economic interest. As for Covid-19, antibody tests provide a veritable supplement to the current testing approach. While the antibody tests only detect the immune system's response to the virus' presence, they are cheap and yield results fast. This makes them ideal for the use of random sampling in the population. The methods illustrated in detail in chapter 2 assist in planning and interpreting the random samples. In short, our approach uses a Maximum-Likelihood-Estimator (ML-estimator) and corrects for the tests' sensitivity and specificity. This helps the researcher to predict the statistical properties of the estimator for a planned random sample and to interpret its results. Tu further assist this process, we provide three calculators programmed with Microsoft Excel that calculate the standard deviation of the estimator for a planned sample or, in reverse, dictate the needed sample size for a required precision of the estimator. Chapter 3 introduces these calculators. Lastly, we apply our considerations to an epidemiological model of infectious diseases in chapter 4. We show for two epidemic scenarios that random sampling methods provide significantly better insights into the true rate of infection in the population than current approaches do.

2 Concepts and Methods

We estimate the true rate of infection in the entire population from the rate of positive test results in a random sample taking into account the known sensitivity and specificity of the underlying test via a maximum likelihood estimator.

The corresponding model is as given: Let us define $p = P(y = 1)$ as the probability of being infected and $q = P(x = 1)$ the probability that a test

is positive. The sensitivity of a test is the conditional probability that the test is positive given that the person is infected $s = P(x = 1|y = 1)$ and the specificity of the test is conditional probability that the test is negative given the person is not infected $z = P(x = 0|y = 0)$.

The test result follows a Bernoulli distribution:

$$P(x) = q^x(1-q)^{(1-x)}. \tag{1}$$

Since the aim is the estimation of p, we reparameterize the distribution by stating that a test can be positive when an infected person gets a right test results and a non-infected gets a wrong test result $q = s \cdot p + (1-z) \cdot (1-p)$. A test can be negative when an infected person gets a wrong test decision and a non-infected gets a right test result $(1-q) = (1-s) \cdot p + z \cdot (1-p)$. The distribution can be written as follows:

$$P(x) = [s \cdot p + (1-z) \cdot (1-p)]^x[(1-s) \cdot p + z \cdot (1-p)]^(1-x). \tag{2}$$

Given a sample of size n the corresponding likelihood takes the following form:

$$L(p|x_1, ..., x_n) = \prod_{i=1}^{n}[s \cdot p + (1-z) \cdot (1-p)]^{x_i}[(1-s) \cdot p + z \cdot (1-p)]^{(1-x_i)}. \tag{3}$$

The loglikelihood is

$$l(p|x_1, ..., x_n) = \sum_{i=1}^{n} x_i \log[s \cdot p + (1-z) \cdot (1-p)] + \sum_{i=1}^{n}(1-x_i)\log[(1-s) \cdot p + z \cdot (1-p)], \tag{4}$$

which can be rewritten as

$$l(p|x_1,...,x_n) = n_1 \log[s \cdot p + (1-z) \cdot (1-p)] + (n-n_1)\log[(1-s) \cdot p + z \cdot (1-p)], \quad (5)$$

where n_1 denotes the number of positive tests in the sample.

The score function of the log likelihood is the partial derivative with respect to p:

$$\frac{\partial l}{\partial p} = n_1 \frac{s+z-1}{s \cdot p + (1-z) \cdot (1-p)} + (n-n_1)\frac{1-s-z}{(1-s) \cdot p + z \cdot (1-p)}. \quad (6)$$

Setting the score function to zero and solving for $\hat{p}$ gives the maximum likelihood estimator, which can be denoted as

$$\hat{p} = \frac{n_1 \cdot s - (n-n_1)(1-s)}{n(s+z-1)}. \quad (7)$$

The variance of the ML estimator can be derived as the inverse of n times the the outer product of the score Sc. That is:

$$1/var(\hat{p}) = n_1 \left(\frac{s+z-1}{s \cdot p + (1-z) \cdot (1-p)} \right)^2 + (n-n_1) \left(\frac{1-s-z}{(1-s) \cdot p + z \cdot (1-p)} \right)^2. \quad (8)$$

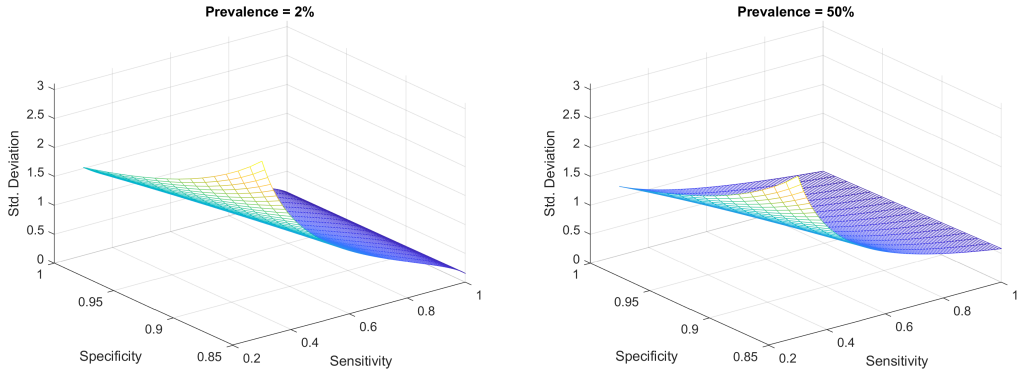
Let us define w_1 the share of positive tests. Thus, $n_1 = w_1 \cdot n$ and $n - n_1 = n(1-w_1)$. Thus, the inverse of the variance can be written as:

$$1/var(\hat{p}) = n \left[w_1 \left(\frac{s+z-1}{s \cdot p + (1-z) \cdot (1-p)} \right)^2 + (1-w_1) \left(\frac{1-s-z}{(1-s) \cdot p + z \cdot (1-p)} \right)^2 \right]. \quad (9)$$

This equation relates the sample size and the probabilities of being infected

Figure 1: The accuracy of the ML-estimator depends on the test quality

(a) Example 1: 2% infection rate (b) Example 2: 50% infection rate



and testing positive to the standard deviation of the estimator. Calculator 1 automates this procedure and yields the standard deviation given the various inputs.

Also, the equation can be solved for n quite easily. This allows the researcher to anticipate the needed size of a random sample given a desired precision expressed in standard deviation. Calculators 2 and 3 make this approach easily accessible.

At this point, we stress the importance of the test's sensitivity specificity for the standard deviation in above calculations.

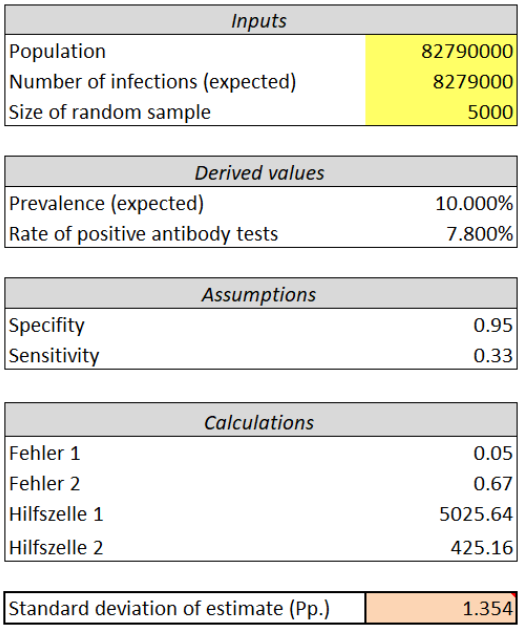
Figure 1 shows the calculated standard deviation of the ML-estimator for a sample size of 10,000 on the vertical axis. The standard deviation does depend on both the sensitivity and specificity, as well as the true rate of infection that is to be determined.

3 Calculators

3.1 Calculator 1

The first calculator produces the standard deviation of the estimator. The only inputs required from the user are the total population in question, the expected number of infected people and the planned size of the random sample.

Figure 2: Calculator 1 - Approximating the Standard Deviation



Inputs	
Population	82790000
Number of infections (expected)	8279000
Size of random sample	5000

Derived values	
Prevalence (expected)	10.000%
Rate of positive antibody tests	7.800%

Assumptions	
Specifity	0.95
Sensitivity	0.33

Calculations	
Fehler 1	0.05
Fehler 2	0.67
Hilfszelle 1	5025.64
Hilfszelle 2	425.16

Standard deviation of estimate (Pp.)	1.354

The population is by default set to the population of Germany. One also needs to insert the expected number of infections. While it is, of course, the ultimate goal of the random sample to calculate the true rate of infection, the accuracy of the estimation tool depends itself on the rate of infection. Therefore, the calculator requires at least an estimate of the number of infected people. The third input is the intended sample size.

Next, the calculator shows two derived values: The expected prevalence is simply the ratio of the number of infected people and the population with which the user has supplied the calculator. Given the expected prevalence one would expect a sample to have positive outcomes at the rate of positive antibody tests. These differ because one has to correct for the sensitivity and the specificity of the tests lower than 1. These qualities of the test are summarized under assumptions. The values in the example of a 0.95 specificity and a 0.33 sensitivity relate to current antibody tests for Sars-CoV-2 and should be updated if the test or the test quality changes. For instance, the rRT-PCR test for Sars-CoV-2 has a much higher sensitivity with which the calculator can be updated.

Figure 3: Calculators 2 and 3 - Approximating the needed sample size

(a) Calculator 2

Inputs	
Population	82790000
Number of infections (expected)	4139500
Standard dev. (Percentage points):	2.5

Derived values	
Prevalence (expected)	5.000%
Rate of positive antibody tests	6.400%

Assumptions	
Specifity	0.95
Sensitivity	0.33

Caclucations	
Testing error 1	0.05
Testing error 2	0.67
Std. dev. (decimal)	0.025
Hilfszelle 1	19.14
Hilfszelle 2	0.09

Min. Required Sample Size	1223

(b) Calculator 3

Inputs	
Population	82790000
Number of infections (expected)	827900

Derived values	
Prevalence (expected)	1.000%
Rate of positive antibody tests	5.280%
Required std. deviation (Pp.)	0.5

Assumptions	
Specifity	0.95
Sensitivity	0.33
Precision factor	0.5

Calculations	
Fehler 1	0.05
Fehler 2	0.67
Std. Abw. (Dezimalzahl)	0.01
Hilfszelle 1	28.12
Hilfszelle 2	0.09

Min. Required Sample Size	25516

The output cell highlighted in orange reports the expected standard deviation in percentage points. In the setting displayed, a rate of infection of 10% can be verified with a standard deviation of ± 1.35 percentage point. The graph to the right of the respective calculator shows the result as a function of different levels of the prevalence. The higher the uncertainty over the expected rate of infection the more advisable it is to consult this sensitivity analysis.

3.2 Calculators 2 and 3

Calculators 2 and 3 calculate the needed sample size in order to estimate the rate of infection with a given precision. The second calculator should be used for an expected prevalence of 5% or above and the third for a prevalence below 5 %. The two tools have in common that they demand the desired standard deviation as input and yield the minimum required sample size as output.

The second calculator is almost identical to calculator 1, except that the cells for the standard deviation and the sample size are exchanged.

The output of calculator 2 should be interpreted as follows: To obtain an estimate for an expected prevalence of 5% with a standard deviation of 2.5 or better, one must at least test 1,223 individuals.

For any expected prevalence below 5% the use of the third calculator is recommended. A more nuanced analysis is necessary because common choices of standard deviations are too large to accurately measure very small rates of infection. A standard deviation of 2.5 from the previous example would be insufficient to differentiate between rates of infections of 1%,0.1% or 0.01%. Following (Naing et al., 2006), calculator 3 internalizes the choice of the needed standard deviation by dividing the expected rate of infection by half. Hence, for an expected rate of infection of 1% the calculator automatically chooses a required standard deviation of 0.5 percentage points. This scaling factor can be set to any desirable value by changing the precision factor in the assumptions cells.

4 Application

In this illustration we model two outbreaks using a canonical model in epidemiology and show that the results of random samples are better suited than currently applied methods to estimate the true rate of infection.

We use the well-known *susceptible-exposed-infected-recovered* model (SEIR) that has been adapted by Peng et al. (2020) and modeled by Cheynet (2020) to reflect a quarantine response by governments. The basic structure of the model is as follows: Individuals advance from a pool of susceptible individuals (S) to those exposed (E), infected (I), quarantined (Q) and recovered (R) or dead (D). The changes of these compartments are given by these derivatives over time:

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\alpha S(t) - \beta \frac{S(t)I(t)}{N_{pop}} \tag{10}$$

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t} = -\gamma E(t) + \beta \frac{S(t)I(t)}{N_{pop}} \tag{11}$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \gamma E(t) - \delta I(t) \tag{12}$$

$$\frac{\mathrm{d}Q(t)}{\mathrm{d}t} = \delta I(t) - \lambda(t)Q(t) - \kappa(t)Q(t) \tag{13}$$

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \lambda(t)Q(t) \tag{14}$$

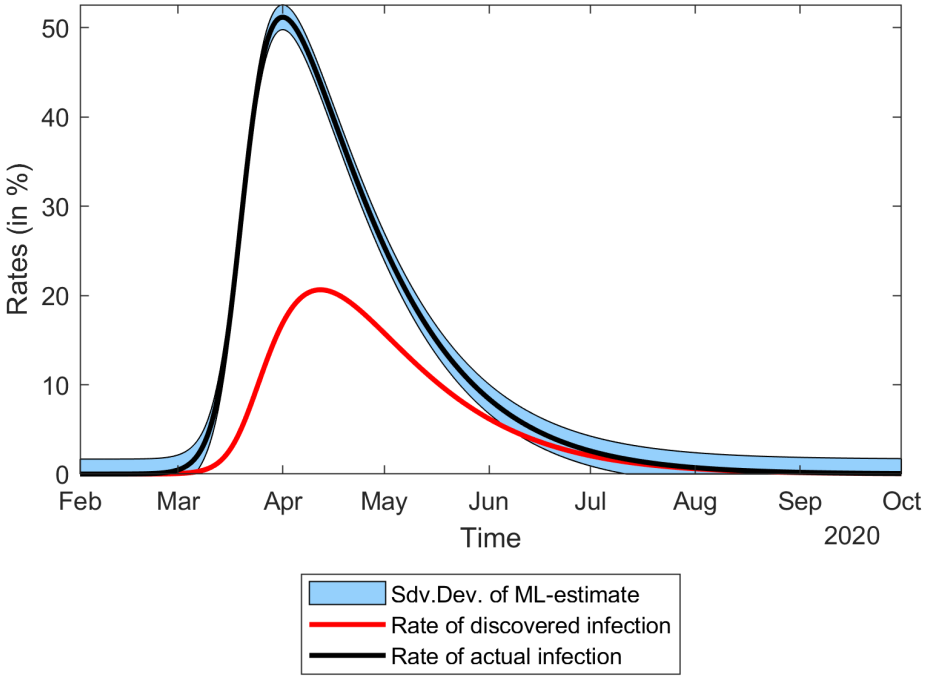
$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = \kappa(t)Q(t) \tag{15}$$

The parameters are the protection rate α, the infection rate β, the inverse of the average incubation period γ, the inverse of the average quarantine time δ, the cure rate λ and the mortality rate κ. In contrast to other SEIR-models this application does not feature immigration, the birth rate or the natural mortality rate to reflect the short-time nature of the outbreaks modeled. The mortality and cure rate of the infection are independent of time for similar reasons. Note that at any point in time, the fraction δ of the infected individuals are discovered to be infected and quarantined. This number of infected and discovered people Q is what is known to the public. In contrast, the true rate of infection is the sum of quarantined infected Q and the undiscovered infected I. Random sampling aims to discover this total rate.

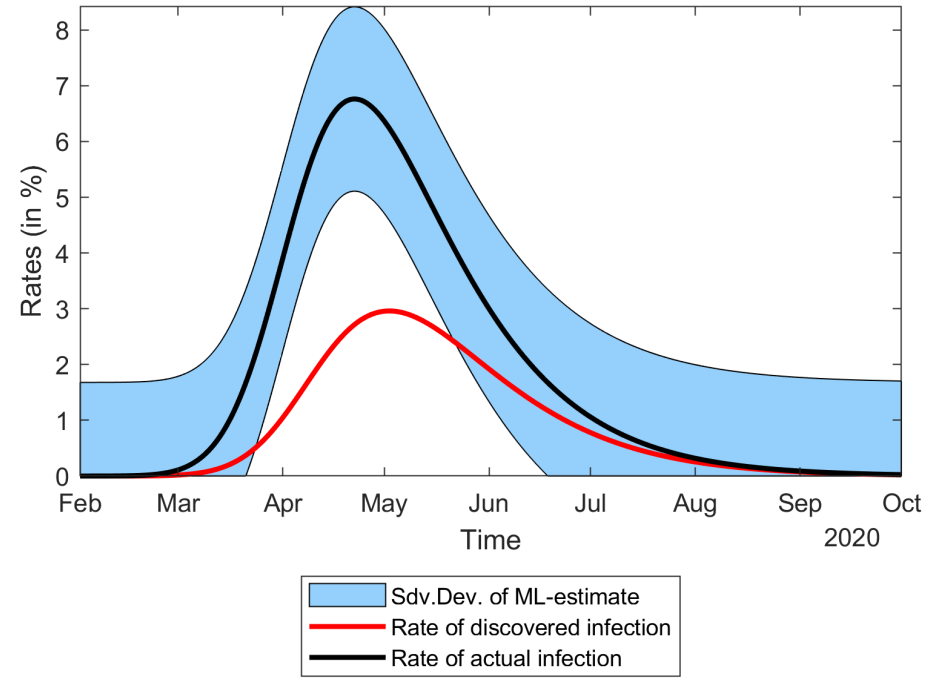
The models are parameterized with a latency period of five days, three weeks in quarantine, an initial population of 83 Million, 100 initial infections occurring on February 1st and and an infection rate of 1. The ML-estimator is produced correcting for a sensitivity of 33% and a specificity of 95%. Scenario (A) shows the dynamics of an outbreak without any successful attempt to slow the spread of the disease (β =0.01). This scenario reaches a herd immunity of 60%. In scenario (B) on the other hand, some measures are in place to reduce the extent of the epidemic and at most 7% of the population are infected

Figure 4: Random sampling reveals undiscovered infections

(a) Scenario (A)



(b) Scenario (B)



at the same time. The public would observe the rate of discovered infection, i.e. individuals that are quarantined, shown as the red line. The true rate of infection instead is shown in black. The ML-estimator from random samples involving 10,000 tests traces this total rate of infection with a standard deviation exemplified by the blue ribbon. These examples clearly show that the ML-estimator would differ from the rate of infection derived from the discovered infections. The scenarios also show that the statistical power of a random sample including 10,000 people is better suited for uncontrolled outbreaks than for more controlled ones.

5 Conclusion

We provide support to assist random sample analyses to determine the true rate of infection in a population. The Maximum-Likelihood estimator provides a way to correct for the medical tests' sensitivity and specificity, which influence the estimator's accuracy. The Excel-calculators allow the researcher to analyze the statistical power of such random samples or to plan the needed sample size. In addition to that, we model two pandemic outbreaks using the SEIR-model and demonstrate that random samples will improve the public's knowledge of the true rate of infection. We do highlight that the statistical power of random samples decreases the more rare a disease is in a population. An estimate of the broad progress of a pandemic outbreak is nevertheless feasible.

References

Andrew Burns, Dominique Van der Mensbrugghe, and Hans Timmer. *Evaluating the economic consequences of avian influenza*. World Bank, 2006.

E. Cheynet. Generalized seir epidemic model (fitting and computation), 2020. URL https://www.github.com/ECheynet/SEIR.

Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine*, 2020.

L Naing, T Winn, and BN Rusli. Practical issues in calculating the sample size for prevalence studies. *Archives of orofacial Sciences*, 1:9–14, 2006.

Liangrong Peng, Wuyue Yang, Dongyan Zhang, Changjing Zhuge, and Liu Hong. Epidemic analysis of covid-19 in china by dynamical modeling. *arXiv preprint arXiv:2002.06563*, 2020.