A SAMPLE DESIGN FOR RELIABLE ESTIMATES OF THE SARS-CoV-2 EPIDEMIC’S PARAMETERS

Calling for a protocol using panel data

1. BACKGROUND AND SCOPE

The urgent need to control the spread of SARS-CoV-2 requires an accurate evaluation of the data sources used for the estimation of the epidemic’s core parameters. Such an approach would be of the utmost importance to forecast and run a simulation of the likely evolutions of the disease; these scenarios are the essential basis for the policy-making decisions directed to an effective healthcare response.

More specifically, in this first phase of big uncertainty concerning the dynamics of a mostly unknown and unfamiliar phenomenon, it is crucial to be able to measure it as accurately as possible. While some degree of uncertainty is inherent in any statistical modelling, the level of inaccuracy in monitoring the development of the situation can and must be kept under control.
Indeed, until now, data in Italy have been collected as a “convenience” sampling, testing mostly cases which display symptoms without following a proper “sample design”. However, as it is well known, convenience sampling can produce biased estimates and do not allow the possibility to attach a given predetermined level of accuracy to them. In this respect, for instance, several studies have clearly shown that the available data strongly underestimate the number of infected people in that they are unable to capture the asymptomatic cases with a possible overestimation of the lethality rate. On the other hand, the recent broad-based data collection of medical swabs carried out in some Italian regions does not constitute a probabilistic sample either. Indeed, for instance, systematically collecting observations in the vicinity of supermarkets leads to an over-inclusion of healthy and/or asymptomatic people in the sample, and to a systematic exclusion of those who (either because they are manifesting symptoms or in any case feel weak) have chosen to stay confined at home.

In order to define proper health care policies, it is important for policy makers to have a clear understanding of the dynamics of the situation in progress in order to take appropriate measures. Similarly, it is of paramount importance for the population to have a full understanding of the situation to guide their individual behaviors.

In such a situation, it is essential to set-up a system of data collection which is able to grant unbiased estimates and to allow statistically significant comparisons through time and across different geographic areas, by taking into account the different economic, demographic, social, environmental and cultural contexts.

We cannot simply limit ourselves to an accurate measurement of the number of infected people, of those that are hospitalized, of their recoveries and of the results of tests. It is, indeed, essential to shed light on the still unknown characteristics of the epidemic which might favor or prevent infection at the individual and family level, and to correctly evaluate the effects of the policies adopted to modify the development of the epidemic. This can only be achieved if epidemiologists, virologists, administrators of healthcare institutions work in conjunction with experts in statistical modeling and forecasting, and experts in the evaluation of public policies.

In order to obtain reliable and useful estimates, we deem it essential to design and implement a rigorous sampling protocol in order to correctly infer the impact of Covid19 on the entire Italian population, using medical swabs or alternative testing procedures such as blood testing and DNA mapping.

This approach would allow us to give more value to the precious information that has already been collected by the healthcare institutions. Indeed, integrating the existing data into a more comprehensive framework with ad-hoc collected data following a precise sampling design would provide a more accurate estimation of the variables of interest for time and selected spatial domains, so as to be able to effectively monitor the spread of the disease and to evaluate the effectiveness of political actions and interventions.

In order to be able to build up a coherent database (which is both relevant and unbiased), it is necessary to integrate different resources: those collected by the administrative institution when receiving and treating individuals that have turned to the healthcare system, and the statistical information collected to accurately measure the diffusion of the infection and those obtained through new sources – such as mobile phones- for tracking the movement of persons and their contacts..

This is a big challenge from the methodological, technological and organizational point of view. A challenge to which the community of statisticians, as well as the officials of public statistics, may offer a useful contribution.

* * * * *

Here below, we propose an observational protocol for the estimation of the people infected by SARS-CoV-2 (with a reliability which is measurable over time and space) according to the various categories of severity.

Starting with a population where it has been ascertained that individuals are infected (*verified*), the aim is to estimate the population which is infected, but has not yet been diagnosed (*asymptomatic*).

For this purpose , the individuals will be pre-classified into two sub-groups which we will refer to as: *target A* and *target B*.

Target A. This is the sub-group consisting of the individuals in which a state of infection has been verified by resulting positive to COVID19 (who could be either hospitalized or in compulsory quarantine) and all the persons who had contact with them in the previous 14 days; therefore, this is the group of individuals who are foreseen to be infected and not only those for whom the infection has already been ascertained. They will represent, therefore, both the *apparent* and *latent* dimensions of the epidemic.

Target B. This second sub-group consists of all the people who have not been in contact with persons in *Target A*. Thus, target B contains healthy persons (i.e not tested), those in which the infection is considered *latent* and those who are still in a phase of incubation where the symptoms can become evident in a future moment of time, in the course of a maximum of 14 days.

Estimates relative to the two sub-groups may be obtained on the basis of a continuous observation in time and following two distinct methodologies, both based on what is known as *indirect sampling*, the same that is used for sampling and estimation of rare and elusive populations (e. g. animal populations).

Here below are reported some ideas as to how to carry out the observation of particular samples for each of these two groups.

2. SAMPLING PROCEDURE FOR TARGET GROUP A

The procedure for sampling target group A in continuous time develops in the following phases:

A1. selection of a sample of infected persons in a predefined unit of time (day, week, other).

A2. Reconstruction of all contacts with others going back 14 days for people selected in the sample point A1; to simplify this phase, applications on a smartphone can be used to track these individuals.

A3. Testing of a pre-defined number of persons who had contacts with the sample selected in point A1.

The sampling of phase A1 must be done taking temporal and spatial dimensions into due consideration. This will involve:

- ✔ *Spatial sampling*. This is done by selecting a sample panel of healthcare institutions (on the basis of their size or other characteristics). The use of spatial sampling techniques allows us to obtain a sample which minimizes the sample size at any given level of accuracy while satisfying some optimality conditions.
- ✔ *Time sampling*. Using the healthcare institutions selected for the spatial sampling, a sample of infected persons will be chosen in order to reconstruct their contacts. This sampling mechanism should be continuous in time and could be carried out within predefined intervals of time, determined by access to the health institution (admission, other). For example, a sample of infected people could be selected among those who had access to the healthcare institution during the course of a two-hour time period.

In order to obtain statistical estimates of the number of infected persons in a given *spatial* (the whole national territory or specific geographic area such as, for example, a region) and *temporal* domain (week/day), it will be sufficient to select about 1,000 individuals to test among the contacts of the infected set of persons. This sample size would ensure a reliable estimate with a sampling error less than 5% under the assumption that the proportion of infected people in the target population is about 20%.

Assuming around 25 contacts for every infected person, the people to whom the 1,000 tests should be administered can be identified by choosing:

- ✔ about 200 infected people whose infections have been verified (from which we obtain a number of 5,000 = 200 times 25).
- ✔ testing a total of about 20% of the total contacts specified above.

Remark 1. It is clear that for health and wellbeing reasons, and in order to prevent the spread of the infection, it would be best to examine all infected people. However, from the statistical point of view, in order to obtain estimates of the highest quality regarding the number of infected persons, this is not strictly necessary. From this point of view it is more important to be able to repeat the sample experiment regularly through time, an aim which would be unsustainable with a complete examination of the whole Italian population.

3. SAMPLING PROCEDURE FOR TARGET GROUP B

This sampling mechanism is carried out to observe the complementary set of people not belonging to target group. For this purpose, an independent *panel* of individuals is selected which needs to be monitored continuously over time. To achieve this aim the following phases are in order:

B1. The panel is subjected to a regular test (for example, once a week),

B2. If a member of the panel gets a positive test result, all of his contacts for the past 14 days are reconstructed,

B3. The test is also administered to a sample of these contacts.

The number of persons involved in a *panel* may be about 1,000 (to obtain around 1,200 tests) for a given *territorial* and *temporal* sampling domain, thus guaranteeing a reliable estimation with a sampling error of less than 5% assuming that the proportion of infected people in this target population is around 4%.

The *panel* should be constructed using a two-step methodology so as to be able to select, by using *pre-screening, two sub-groups*.

- ✔ a certain number of individuals who continue to travel (and are therefore more subject to being infected);
- ✔ a number of people with few contacts who fully observe the prescribed quarantine recommendations.

Remark 2. The overlap we could observe between the two samples (the one for target group A and the other for target group B) can be used to estimate the intersection of the two target groups.

The formal description of the sampling and estimation will be released on this website

https://web.uniroma1.it/memotef/sites/default/files/Proposal.pdf in a few days.

Rome, April 2, 2020

Giorgio Alleva, Full Professor of Statistics, Sapienza University of Rome, Former President of the Italian National Statistical Institute (Istat)

Giuseppe Arbia, Full professor of Economic Statistics, Catholic University of the Sacred Heart, Milan, Spatial Data Analysis Expert

Piero Demetrio Falorsi, Former Director of the Methodological Directorate of the Italian National Statistical Institute (Istat), Expert in Sample Designs

Guido Pellegrini, Full Professor of Economic Statistics, Sapienza University of Rome, Expert in Public Policy Evaluation and President of the Statistical Information Guarantee Commission

Alberto Zuliani, Emeritus Professor of Statistics, Former President of the Italian National Statistical Institute (Istat)