ESTIMATING THE NUMBER OF INFECTED PERSONS

YANNICK BARAUD, IVAN NOURDIN, AND GIOVANNI PECCATI

Let n_T be the number of individuals that have been infected within a window of time T (a week or a longer period) and r be the mortality rate, i.e. the probability that an infected person dies. We observe the number N of deaths within a window of time W (say T plus two weeks), and we assume that if an individual who has been infected in the window of time T dies, the death is likely to occur within the observed period of time W with a probability $\beta \in (0,1]$ close to 1. This leads us to model N as a random variable drawn from a binomial distribution with parameters n_T and $r\beta$; that is, $N \sim \mathcal{B}(n_T, r\beta)$. Indeed, mathematically speaking everything happens as if each affected individual tosses an unfair coin (showing head with probability $r\beta$), and dies if the coin falls on head.

Our goal is to estimate the value of n_T from the observation of N and the prior knowledge of r and β (actually we shall see later that the choice $\beta = 1$ might be a reasonable one anyway). Since the expectation of the binomial distribution with parameters n_T and $r\beta$ is $n_T r\beta$, whenever N is close enough to its mean, n_T is of order $N/(r\beta)$ and the latter quantity provides a natural estimation of the unknown parameter n_T. However, N fluctuates around its mean and we would like to take this into account to build a confidence interval for n_T rather than giving a rough estimation of it. Since it is difficult to work directly with the binomial distribution, we shall approximate it with the Gaussian and Poisson ones, each of these approximations will lead to a confidence interval for n_T. For values of r of a few percent and n_T larger than a thousand, both distributions provide a very good approximation of the binomial and consequently both methods can be used to solve the problem. Nevertheless, the special properties of the family of Gaussian distributions (which is translation and scale invariant) will enable us to provide a more accurate confidence interval.

1. The Gaussian approximation

It is classical to approximate the binomial distribution $\mathcal{B}(n,p)$ by the Gaussian distribution $\mathcal{N}(np, np(1-p))$ with the same mean np and the same

Date: April, 21 2020.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 811017.

variance $np(1-p)$. In this section, we shall therefore assume that the distribution of the number of deaths N follows the distribution $\mathcal{N}(n_T r\beta, n_T r\beta(1-r\beta))$. Under such an assumption, the following result holds.

Proposition 1. *Let $\alpha \in (0,1)$, q_α be the $(1-\alpha)$-quantile of a standard Gaussian random variable and*

$$\widehat{z}_\alpha = \frac{(1-r\beta)q_\alpha^2}{2N}.$$

Each of the inequalities below holds true with a probability $1-\alpha$,

$$n_T \leq \frac{N}{r\beta}\left[1+\sqrt{\widehat{z}_\alpha(2+\widehat{z}_\alpha)}+\widehat{z}_\alpha\right] \quad \textit{and} \quad n_T \geq \frac{N}{r\beta}\left[1-\sqrt{\widehat{z}_\alpha(2+\widehat{z}_\alpha)}+\widehat{z}_\alpha\right]$$

In particular, both inequalities are satisfied simultaneously with a probability $1-2\alpha$.

2. The Poisson approximation

When p is small, it is also classical to approximate the binomial distribution $\mathcal{B}(n,p)$ by the Poisson distribution $\mathcal{P}(np)$. In this section we shall therefore assume that the distribution of N is Poisson with parameter $n_T r\beta$. The following result holds.

Proposition 2. *Let $c>0$ be some positive number. Each of the following inequalities holds with a probability at least $1-e^{-c}$,*

$$n_T \geq \frac{N}{r\beta}\left[1-\sqrt{\frac{2c}{N}\left(1+\frac{2c}{9N}\right)}+\frac{2c}{3N}\right]$$

and

$$n_T \leq \frac{N}{r\beta}\left[1+\sqrt{\frac{2c}{N}\left(1+\frac{c}{2N}\right)}+\frac{c}{N}\right].$$

In particular, both inequalities are satisfied simultaneously with a probability at least $1-2e^{-c}$.

3. Example and discussion

If we take $N=75$ (number of deaths from the beginning of the pandemic until now), T the period of time from the beginning of the pandemic until two weeks ago, $\beta=1$ and $r=2\%$, we obtain that with a probability (at least) 90%,

$$3109 \leq n_T \leq 4524 \quad \text{(Gaussian approximation)}$$
$$2786 \leq n_T \leq 4970 \quad \text{(Poisson approximation)}.$$

Some remarks:

- Note that the estimation bounds on n_T are quite sensitive with respect to the value of r (and also that of N since r is small). Underestimating or overestimating r may therefore have great consequences on the bounds we get.

- As already observed, the "true" distribution of the random number N is of course not Gaussian, nor Poisson, but rather binomial with parameter $r\beta n_T$. It is always possible to "go back" to the true distribution of N by using some classical approximation results. Indeed, one can combine Proposition 2 with well-known bounds by Le Cam, Chen, Stein, Daley, Vere-Jones, ... implying that the distribution of N is well approximated by a Poisson distribution with parameter $r\beta n_T$ up to a uniform error (over generic sets) which is not bigger than $0.71 \times r\beta$. So, even if we assume that N has the binomial distribution, the conclusion of Proposition 2 still holds with $(1 - 2e^{-c})$ replaced by $(1 - 2e^{-c}) - 0.71 \times r\beta$. With the parametrization of the previous example, one has $r\beta = 2\%$, $0.71 \times r\beta \sim 1.4\%$ and the estimate $2786 \leq n_T \leq 4970$ holds with a probability of at least $88,6\%$. Similarly, it is known ("Berry-Esseen bound") that the distribution of N can be approximated with a Gaussian distribution of mean $n_T r\beta$ and variance $n_T r\beta(1 - r\beta)$ with a uniform error (over half-lines) which is not bigger than

$$E := \frac{0.5}{\sqrt{n_T}} \left(\frac{(1 - r\beta)^{3/2}}{\sqrt{r\beta}} + \frac{(r\beta)^{3/2}}{\sqrt{1 - r\beta}} \right).$$

  So, even if we assume that N has the binomial distribution, each one of the two inequalities in the statement of Proposition 1 still hold with probability $1 - \alpha/2$ up to an error of maximal magnitude E. With the parametrization of the previous example, one has that $r\beta = 2\%$ and, assuming $n_T \geq 3500$, $E \leq 0.06$, and the probability that $n_T \leq 4524$ is at least 89% (using the fact that the probability of the one-tailed event $n_T \leq 4524$ under the assumptions of Proposition 1 is 95%).
- If we look at the number N of deaths in week W and want to estimate the number of people who were newly infected two weeks ago (say), the previous formulas apply provided that we can properly estimate the probability β that an infected person will not die in week W, meaning that his or her death will occur before or after the week W. Nevertheless, the estimation of β may not be as crucial for the following reason. On the one hand, N includes the number of deaths of those who were not infected two weeks before week W, but on the other hand, it does not include the deaths of these people who were infected two weeks ago and died outside the window of time W. The two numbers of deaths can eventually compensate one

another, meaning that the choice $\beta = 1$ could make sense even in the unfavorable situation where the deaths of some people infected two weeks ago were not observed during week W.

- Note that we could alternatively estimate n_T from the observation of the number of severe cases in hospitals. In this case r would correspond to the probability of clinical severity and the same bounds as those described in the proposition apply. It is then possible to build another confidence interval I_2 for n_T. Intersecting I_2 with the one obtained previously, say I_1, results in a new confidence interval $I_1 \cap I_2$ the length of which cannot be larger than those of I_1 and I_2. Note that if I_1 and I_2 have a confidence level $1-2e^{-c}$, that of $I_1 \cap I_2$ is however $1-4e^{-c}$.

4. Proofs

4.1. **Proof of Proposition 1.** Let $X \sim \mathcal{N}(0,1)$ be a standard Gaussian random variables. Under our assumption, N has the same distribution as

$$n_T r\beta + \sqrt{n_T r\beta(1-r\beta)}X.$$

Since with a probability $1-\alpha$, $X \geq -q_\alpha$, we obtain the first inequality by solving the inequality

$$N \geq n_T r\beta - \sqrt{n_T r\beta(1-r\beta)}q_\alpha.$$

The second inequality is obtained by arguing similarly using that with a probability $1-\alpha$, $X \leq q_\alpha$.

4.2. **Proof of Proposition 2.** The proof of Proposition 2 relies on the following lemma.

Lemma 1. *Let X be a random variable with Poisson distribution with parameter $\theta > 0$. For all $c > 0$*

$$\mathbb{P}\left[N - \theta > \frac{c}{3}\left[1 + \sqrt{1 + 18c^{-1}\theta}\right]\right] \leq e^{-c} \quad \textit{and} \quad \mathbb{P}\left[N - \theta < -\sqrt{2c\theta}\right] \leq e^{-c}.$$

Proof. Let $z > 0$. The mapping

$$F : \lambda \mapsto \left(e^\lambda - \lambda - 1\right)\theta - \lambda z \quad \text{on } \mathbb{R}_+$$

is minimum for $\lambda^\star(z) = \log(1 + z/\theta)$ and $F(\lambda^\star(z)) = -\theta h\left(z/\theta\right)$ where

$$h(u) = (1+u)\log(1+u) - u \geq \frac{u^2}{2(1+u/3)} \quad \text{for all } u > 0.$$

Hence

$$F(\lambda^\star(z)) = -\theta h\left(z/\theta\right) \leq -\frac{z^2}{2(\theta + z/3)}. \tag{1}$$

The mapping

$$G : \lambda \mapsto \left(e^{\lambda} - \lambda - 1\right)\theta + \lambda z \quad \text{on } \mathbb{R}_-$$

is minimum for $\lambda^\star(-z)$ and $G(\lambda^\star(-z)) = -\theta g\left(z/\theta\right)$ where

$$g(u) = (1-u)\log(1-u) + u \geq \frac{u^2}{2} \quad \text{for all } u > 0.$$

Hence,

$$G(\lambda^\star(-z)) = -\theta g\left(z/\theta\right) \leq -\frac{z^2}{2\theta}. \tag{2}$$

Using that

$$\mathbb{E}\left[e^{\lambda(X-\theta)}\right] = \exp\left[(e^{\lambda} - \lambda - 1)\theta\right] \quad \text{for all } \lambda \in \mathbb{R},$$

we deduce that for all $z > 0$

$$\mathbb{P}\left[N - \theta \geq z\right] \leq \exp\left[F(\lambda^\star(z))\right] \leq \exp\left[-\frac{z^2}{2(\theta + z/3)}\right].$$

In particular, the right-hand side equals e^{-c} if

$$c = \frac{z^2}{2(\theta + z/3)} \iff z = \frac{c}{3}\left[1 + \sqrt{1 + \frac{18\theta}{c}}\right].$$

Similarly,

$$\mathbb{P}\left[X - \theta \leq -z\right] \leq \exp\left[G(\lambda^\star(-z))\right] \leq \exp\left[-\frac{z^2}{2\theta}\right]$$

and the right-hand side equals e^{-c} for $z = \sqrt{2c\theta}$. □

It follows from the lemma that with a probability at least $1 - e^{-c}$,

$$X - \theta \leq \frac{c}{3}\left[1 + \sqrt{1 + 18c^{-1}\theta}\right] \iff \theta \geq X\left[1 - \sqrt{\frac{2c}{X}\left(1 + \frac{2c}{9X}\right)} + \frac{2c}{3X}\right].$$

In the other way, with a probability at least $1 - e^{-c}$

$$X - \theta \geq -\sqrt{2c\theta} \iff \theta \leq X\left[1 + \sqrt{\frac{2c}{X}\left(1 + \frac{c}{2X}\right)} + \frac{c}{X}\right].$$

The results of the proposition follow by applying these bounds with $X = N$ and $\theta = n_T r\beta$.

Department of Mathematics
University of Luxembourg
Maison du nombre
6 avenue de la Fonte
L-4364 Esch-sur-Alzette
Grand Duchy of Luxembourg
Email address: yannick.baraud@uni.lu
Email address: ivan.nourdin@uni.lu
Email address: Giovanni.Peccati@uni.lu