

# Checking individuals and sampling populations with imperfect tests

Giulio D'Agostini<sup>1</sup> and Alfredo Esposito<sup>2</sup>

## Abstract

In the last months, due to the emergency of Covid-19, questions related to the fact of belonging or not to a particular class of individuals ('infected or not infected'), after being tagged as 'positive' or 'negative' by a test, have never been so popular. Similarly, there has been strong interest in estimating the proportion of a population expected to hold a given characteristics ('having or having had the virus'). Taking the cue from the many related discussions on the media, in addition to those to which we took part, we analyze these questions from a probabilistic perspective ('Bayesian'), considering several effects that play a role in evaluating the probabilities of interest. The resulting paper, written with didactic intent, is rather general and not strictly related to pandemics: the basic ideas of Bayesian inference are introduced and the uncertainties on the performances of the tests are treated using the metrological concepts of 'systematics', and are propagated into the quantities of interest following the rules of probability theory; the separation of 'statistical' and 'systematic' contributions to the uncertainty on the inferred proportion of infectees allows to optimize the sample size; the role of 'priors', often overlooked, is stressed, however recommending the use of 'flat priors', since the resulting posterior distribution can be 'reshaped' by an 'informative prior' in a later step; details on the calculations are given, also deriving useful approximated formulae, the tough work being however done with the help of direct Monte Carlo simulations and Markov Chain Monte Carlo, implemented in R and JAGS (relevant code provided in appendix).

*"Grown-ups like numbers"*  
(The Little Prince)

*"The theory of probabilities is basically  
just common sense reduced to calculus"*  
(Laplace)

*"All models are wrong, but some are useful"*  
(G. Box)

---

<sup>1</sup>Università "La Sapienza" and INFN, Roma, Italia, giulio.dagostini@roma1.infn.it

<sup>2</sup>Retired, alfespo@yahoo.it

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Rough reasoning based on expectations</b>	<b>7</b>
2.1	Setting up the problem . . . . .	7
2.2	Fraction of sampled positives being really infected or not . . . . .	8
2.3	Fraction of infectees in the positive sub-sample . . . . .	9
2.4	Estimating the proportion of infectees in the population . . . . .	11
2.5	Moving to probabilistic considerations . . . . .	12
2.6	Summing up . . . . .	13
<b>3</b>	<b>Probability of infected in the light of the available information</b>	<b>13</b>
3.1	Bayes' rule at work . . . . .	14
3.2	Initial odds, final odds and Bayes' factor . . . . .	17
3.3	What do we learn by a second test? . . . . .	19
<b>4</b>	<b>Uncertainty about <math>\pi_1</math> and <math>\pi_2</math></b>	<b>20</b>
4.1	From $P(n_{P_I}   n_I, \pi_1)$ to $f(\pi_1   n_{P_I}, n_I)$ : Bayes' rule applied to 'numbers' . . . . .	20
4.2	Conjugate priors . . . . .	22
4.3	Expected value or most probable value of $\pi_1$ and $\pi_2$ ? . . . . .	25
4.4	Effect of the uncertainties on $\pi_1$ and $\pi_2$ on the probabilities of interest . . . . .	26
4.5	Adding also the uncertainty about $p$ . . . . .	29
4.6	Uncertainty about $P(\text{Inf}   \text{Pos})$ and $P(\text{NoInf}   \text{Neg})$ ? . . . . .	30
<b>5</b>	<b>Predicting the number of positives resulting from testing a sample</b>	<b>31</b>
5.1	Expected number of positives and its standard uncertainty . . . . .	31
5.2	Taking into account the uncertainty on $\pi_1$ and $\pi_2$ . . . . .	34
5.2.1	Approximated formulae . . . . .	36
5.3	General considerations on the approximated evaluation of $\sigma(n_P)$ . . . . .	39
5.3.1	Contribution of the uncertainty on $p_s$ due to sampling . . . . .	40
<b>6</b>	<b>Sampling a population</b>	<b>41</b>
6.1	Proportion of infected individuals in the random sample . . . . .	42
6.2	Expected number of positives assuming exact values of $\pi_1$ and $\pi_2$ . . . . .	44
6.2.1	Approximated results . . . . .	44
6.3	Detailed study of the four contributions to $\sigma(f_P)$ . . . . .	46
6.4	Statistical and systematic contributions to $\sigma(f_P)$ . . . . .	50

<b>7</b>	<b>Measurability of <math>p</math></b>	<b>56</b>
7.1	Probabilistic model . . . . .	56
7.2	Monte Carlo estimates of $f(n_P)$ and $f(f_P)$ . . . . .	57
7.2.1	Using the R random number generators . . . . .	58
7.2.2	Using JAGS . . . . .	60
7.2.3	Further check of the approximated formulae . . . . .	64
7.3	Resolution power . . . . .	64
7.4	Predicting fractions of positives sampling two populations . . . . .	66
<b>8</b>	<b>Inferring <math>p</math> from the observed number of positives in the sample</b>	<b>68</b>
8.1	From the general problem to its implementation in JAGS . . . . .	68
8.2	Inferring $p$ and $n_I$ with our ‘standard parameters’ . . . . .	70
8.3	Dependence on our knowledge concerning $\pi_1$ and $\pi_2$ . . . . .	72
8.4	Quality of the inference as a function of $n_s$ and $f_P$ . . . . .	74
8.5	Updated $f(\pi_1)$ and $f(\pi_2)$ in the case of ‘anomalous’ number of positives . . . . .	76
8.6	Inferring the proportions of infectees in two different populations . . . . .	78
8.7	Which priors? . . . . .	79
8.7.1	Symmetric role of prior and ‘integrated likelihood’ . . . . .	80
8.7.2	Some examples . . . . .	81
8.7.3	Some approximated rules . . . . .	83
<b>9</b>	<b>Exact evaluation of <math>f(p)</math></b>	<b>85</b>
9.1	Setting up the problem . . . . .	85
9.2	Normalization factor and other moments of interest . . . . .	87
9.3	Result and comparison with JAGS . . . . .	88
9.4	More remarks on the role of priors . . . . .	91
<b>10</b>	<b>Conclusions</b>	<b>93</b>
	<b>References</b> . . . . .	<b>96</b>
	<b>Appendix A – Some remarks on ‘Bayes’ formulae’</b> . . . . .	<b>99</b>
	<b>Appendix B – R and JAGS code</b> . . . . .	<b>103</b>

# 1 Introduction

The Covid-19 outbreak of these months raised a new interest in data analysis, especially among lay people, for long locked down and really flooded by a tidal wave of numbers, whose meaning has often been pretty unclear, including that of the body counting, which should be in principle the easiest to assess. As practically anyone