

Not an exponential curve fit: a data analysis exercise on COVID-19 Italian data trying to estimate the number of infected people

How many people are really infected? Nobody knows. The number of asymptomatic people is high, but nobody has a definitive number. Mostly because any estimate would only be accurate on a small region since the number of tests carried out differ significantly from place to place. Knowing the number of infected is useful for society as many cured people would allow to re-open Europe, at least partially. Since our network is aiming to help society using particle physics tools, I will try to give an estimate based on some of the data available from Italy. Will my estimate be accurate? Probably not, but even if they only help someone to better understand the complex nature of analysing COVID data, I think it is worth sharing them.

Let me start by saying that this is not a quantitative analysis, I do not have the background to do that. I will make some assumptions which should be correct enough for a qualitative analysis and give an approximate result which should be confirmed by experts.

Review of existing estimate

Since I started developing this method, two estimates were produced in Italy. One is described in a [preprint](#), the second is a survey from [Doxa](#). The former uses the Case Fatality Ratio (CFR), defined as the probability to die if infected, in small countries (0.2%), on the Diamond Princess (1%) or in Italian hospital (10%). I find these numbers not very precise as the early cases in small countries like Island, Luxemburg was imported from people travelling. This is usually a relatively young population which we know has a lower fatality rate than the average, so I find this a bit optimistic. The estimate using the cruise ship Diamond Princess is based on the 7 deaths out of more than 700 infected, so the statistical uncertainty is large. Furthermore, the population on a cruise ship probably does not reproduce the demographic of a country. Getting a good age distribution is crucial given the much higher mortality for the older age groups; this is the problem of using the hospital population which is significantly skewed toward the older population. The Doxa analysis is based on 1000 people spread across Italy and counting those describing symptoms associated with the virus. There are several assumptions, but the main problem is that 1000 people is not enough to sample all regions and age groups accurately. In general, all methods try to extrapolate to the whole Italian population which, in my opinion, the biggest problem as the diffusion depends strongly on geography. About half of all Italian cases are in Lombardia so a more granular approach is needed.

The data

Moving to the data, the most granular information available is in the ISS (Istituto Superiore della Sanita') [appendice con dettaglio regionale](#) (appendix with region-by-region information). This appendix is published every week and the latest available data is from the 16th of April. The number of infected is provided for each province but the demographic curves of the infected and the number of deaths are only presented for the regions. It would be great if ISS could share this additional data so that the predictions could be improved.

I also want to make a consideration on the quality of this data. A lot of people on the internet question the validity and the usefulness of all these numbers. As in all data analysis, there is no wrong data, at most, there is bad data and it the job of the analyser to understand it. In general, domain expertise would allow doing this correctly, but in this case, the confusion is generated

by the attempt to extrapolate or aggregate to national level what should be analysed as a localised problem. This is what I will try in my analysis.

The analysis

The figure that intrigued me and led me to start this analysis is the demographic curve of infected people in Italy (an example is provided [here](#)). This is very different from the demographic curve of the whole Italian population ([link](#)). If we could find a segment of the population where most of the infected are identified, we could get a first approximation of the all infected population. I think such a group exists and are the males over 80. I base this assumption on the fact that 40% of the infected male over 80 died because of the virus. Given such high mortality, it is unlikely many will be asymptomatic like most under 40. Since most infected people in this group will be very symptomatic, most of them should be tested and identified. There are several problems with this assumption, for example, people may die before being tested and a small fraction may have very mild or no symptoms at all. Hence, any result based on this assumption is a lower limit. The rate of asymptomatic in this age group could be easily measured by a dedicated study monitoring this group in a region or province; since this is a constant number it could then be applied to the whole country. Indeed, this information could be extracted by the available tests carried out in hospices. I selected male because an important factor in this pandemic is the sex of the infected; in the over 80 population, a similar number of male and female is infected although the population is dominated by women. I do not think there is a substantial difference in the lifestyle of the two sexes, so both should be infected at the same rate. My assumption here is that women simply get milder symptoms and therefore are less tested but will have a similar infection rate. This is supported by the fact that the mortality rate for women is only 20%; it looks like women are more resistant, so could be also more asymptomatic. Since for people over 90, the women-man ratio in the population and in the infected is significantly different from the 80-90, I will focus on the 80-90 population only (all demographic curves use 10 years ranges, so this simplify some of the data extraction too). The first hypothesis is that at least as many people in each region are infected as the 80-90 age group; this hypothesis is denoted IP1 from now on. This assumption is supported by a recent [study](#) done on the completely locked-down town of Vo, in Veneto. The 80-90 age group had an infection rate comparable to the whole population; it must also be noted that only 5 cases were identified, so there is large statistical uncertainty in this confirming result. One side note about this study; it showed that children do not get the infection, so the closing school did not make a big impact and should be among the first measured to be lifted as already done in Denmark, Norway and, soon, in Germany.

In normal conditions, people in their 80s are not the most socially active group around. So, I would expect a significantly higher transmission and infection rate in younger age groups. As a physicist, what I find interesting about this estimate, is that it is easy to correct because it does not require to know the absolute infection rate of the virus, but the relative infection rate between age groups! This could be derived from studies like the one cited above done on the population of Vo. In particle physics, we measure a lot of ratios because we get a lot of effects that cancel out and I hope to achieve the same effect in my study. I had a look at publications in this area, which would be based in non-lockdown conditions, and I could only find this [paper](#). In summary, people below 50 are twice more likely to get infected than 65+. This factor two may not be the most appropriate for a lockdown scenario, it may be smaller (if everyone is really isolated at home as was the case of the study on Vo mentioned above where the factor was 1) or higher if a significant fraction of the younger working population continue to interact (50% of companies are still open as deemed essential) while the older population practice tight

isolation. It is possible to get a rough estimate this factor from the Vo study as they measured the rate of infection at the beginning of the lockdown (so those cases would have been caused by normal circulation of people) and after two weeks of total lockdown. Unfortunately, the new cases were so few that any extrapolation based on the ratio between age group would be meaningless (error bigger than 100%). Comparing the age-inclusive rate has a smaller (but still large) statistical uncertainty; 0.3% during lockdown compared to 2.6% before. Taking the mid-point for the working group would give 1.5%. The ratio between locked down older generation over the working population would be 5 to 1, higher than the factor two mentioned before. When extrapolating to the total population this value must be reduced as the school-age population is in lockdown and those working would probably take measure to reduce infection. Combining all these effects a rough estimate could be that the whole population is only twice as infected as the old; I welcome ideas to improve this estimate as I know it is not great. I will denote this as IP2 from now on.

Let me stress that my estimates do not use the mortality rate at all and can rely on the bigger numbers of identified infected people to extrapolate to the actual number of infected, so statistical errors are smaller and the estimate more accurate. The results are summarised in Table 1. I decided to include some of the intermediated steps for completeness. The only man-women division based on demographic that I could find was on a national scale ([here](#)), indicating that 49.4% of the cases in the 80-90 group were men. I have no reason to suspect that this ratio should be significantly different in any Italian region, so I will assume that this is true in all regions. This is data available to the ISS, so it would be great if they could provide it to make the prediction more precise. The column “Fraction of infected All [%]” is the fraction of infected in each region; the large differences are caused by the different level of spread but also by the significant difference in the number tests performed by each region. Many people complain that COVID data does not make sense citing these numbers; actually, the data is correct but must be understood, for example by avoiding such easy (but wrong) comparisons or by aggregating nonhomogeneous data. By analysing each region independently, I avoid this problem and I find that data make more sense. I provide the summary for Italy and Italy without Lombardia only to stress this point. The latter is also useful to show that Italy is doing much better, especially in the number of deaths per inhabitants, is if the worse affected region is removed from the national calculation.

Table 1 Summary of data and estimates for all Italian regions

Region	Population	Population Male 80-90	Infected	Infected 80-90	Deaths	Infected Male 80-90	Fraction infected Male 80-90 [%] (IP1)	Fraction of infected All [%]	Ratio 80-90/All	IP2 [%]
Lombardia	10060574	225913	54801	9505	9731	4692	2.08	0.54	3.81	4.15
Lazio	5879082	127955	4238	514	230	254	0.20	0.07	2.75	0.40
Campania	5801692	97256	3002	245	122	121	0.12	0.05	2.40	0.25
Sicilia	4999891	104554	1782	181	105	89	0.09	0.04	2.40	0.17
Veneto	4905854	109628	12940	1887	756	932	0.85	0.26	3.22	1.70
Emilia-Romagna	4459477	116600	18188	2926	2221	1444	1.24	0.41	3.04	2.48
Piemonte	4356406	118361	12392	2211	1210	1091	0.92	0.28	3.24	1.84
Puglia	4029053	87511	2714	352	224	174	0.20	0.07	2.95	0.40
Toscana	3729641	101396	5219	745	269	368	0.36	0.14	2.59	0.73
Calabria	1947131	43902	785	71	46	35	0.08	0.04	1.98	0.16
Sardegna	1639591	36589	903	150	57	74	0.20	0.06	3.67	0.40
Liguria	1550640	47816	2899	564	510	278	0.58	0.19	3.11	1.16
Marche	1525271	42068	4755	820	361	405	0.96	0.31	3.09	1.92
Abruzzo	1311580	33589	1862	189	33	93	0.28	0.14	1.96	0.56
Friuli Venezia Giulia	1215220	30861	1987	254	172	125	0.41	0.16	2.48	0.81
Trentino-Alto Adige	1072276	21815	5194	740	439	365	1.67	0.48	3.46	3.35
Umbria	882015	24615	1194	90	50	44	0.18	0.14	1.33	0.36
Basilicata	562869	14435	149	10	16	5	0.03	0.03	1.29	0.07
Molise	305617	8157	232	27	13	13	0.16	0.08	2.15	0.33
Valle d'Aosta	125666	2932	874	155	89	77	2.61	0.70	3.75	5.22
Italy	60359546	1395953	136110	21636	16654	10681	0.77	0.23	3.39	1.53

Italy w/out Lombardy	50298972	1170040	81309	12131	6923	5989	0.51	0.16	3.17	1.02
----------------------	----------	---------	-------	-------	------	------	------	------	------	------

Table 2 Comparison of provinces in Trentino-Alto Adige

Region	Population	Population Male 80-90	Infected	Infected 80-90	Deaths	Infected Male 80-90	Fraction infected Male 80-90 [%] (IP1)	Fraction of infected All [%]	Ratio 80-90/All	IP2 [%]
Trentino-Alto Adige	1072276	21815	5194	740	439	365	1.67	0.48	3.46	3.35
Bolzano	531178	10487	1868	270	184	133	1.27	0.35	3.61	2.54
Trento	541098	11328	3326	470	255	232	2.05	0.61	3.33	4.10

To further prove the need for a province-level granularity of the data, I analysed the two provinces of Trentino Alto Adige since the ISS provide all the details for both. The results are shown in Table 2; Trento, a province closer to the epicentre of the pandemic, has twice as many cases as Bolzano. Using the region data only show the average between the two that is not an accurate description of the spread of the virus.

Study on CFR, a digression

The CFR has been another source of confusion that led people to claim that the available data is wrong or useless. Again, the data is correct and simply should not be compared because of the definition of CFR includes a value that it not comparable between regions. Let me go back to the definition; CFR is the fraction of deaths over infected. While the numerator (the number of deaths) is a well-defined number which is common to all regions, the denominator (the number of infected) suffers from the different number of tests per inhabitants performed in the various regions. For example, Veneto has performed as many tests as Lombardia but has a population that is less than half. This resulted in more people being identified as positive, particularly in lower age groups. This resulted in Veneto having an average age in the infected of 58 while Lombardia average age for the infected in 65.

This problem could be avoided if the demographic of the deceased could be provided as it is done for the infected. With this breakdown, it would be possible to use the 80-90 age control group to compare the CFR between regions; again, I would assume that all regions are identifying all (or at least the majority) of the infected in this age group.

The 80-90 age group can also provide an evaluation of the performance of the regionalised health system in Italy by counting the fraction of deaths in this group. The guidelines to treat patients in case of shortage of ICS places state that priority should be given to the healthier and younger patients that have a higher success rate in surviving. Therefore, a significantly higher number of deaths my control group would be evidence of doctors having to make difficult decisions due to lack of resources.

Unfortunately, the data to carry out this study is not available as only the national breakdown of death in age groups is provided. What can be done is to divide the number of deaths in this age group proportionately to the deaths in each region. This approximation would hide some of the excesses that the study aims to find as the deaths are equally distributed, so any excess found will be underestimated. I already calculated the number of infected in the 80-90 age group which are used as the denominator for the calculation of the fatality. The results for all regions are shown in Table 3. The CFR in Lombardia is the highest of all Italy and is almost twice the average in the rest of Italy. It is also interesting to notice that Liguria has a very high CFR too, something that has not been picked by the media (probably due to the bigger numbers in Lombardia). Emilia-Romagna, the second most affected region also show a higher rate than the average, it would be interesting to have the breakdown by provinces to see if the rate would get worse in the provinces closer to Lombardia which are more affected by the virus. These numbers do not definitively prove that medics had to make difficult choices but hint in this direction. A definitive answer could be given if ISS would provide the additional information mentioned above (the breakdown of deaths by age in each region and provinces).

Table 3 Estimated number of deaths and CFR for the 80-90 age group in all regions

Region	Dead 80-90	CFR 80-90 [%]
Lombardia	4594	39.19
Lazio	105	15.42
Campania	59	19.90
Sicilia	58	25.16
Veneto	396	17.60
Emilia-Romagna	1120	32.40
Piemonte	615	20.31
Puglia	119	25.79
Toscana	151	15.86
Calabria	21	27.08
Sardegna	33	16.80
Liguria	214	34.56
Marche	172	17.23
Abruzzo	15	6.55
Friuli Venezia Giulia	84	29.81
Trentino-Alto Adige	220	23.95
Umbria	23	25.44
Basilicata	8	77.05
Molise	6	22.27
Valle d'Aosta	55	31.59
Italy	8070	30.22
Italy without Lombardy	3476	23.20

A hope from the north of Europe?

The fraction of infected evaluated with IP1 and IP2 are rather low and, even in the most affected regions, the numbers are far from those needed for herd immunity. However, these are rather conservative values and the infection may be more widespread.

Some hope is coming from Sweden than just announced that some regions of the country already show sign of herd immunity and should reach the required levels of infected (more than 60%?) in about a month. While Sweden as been described as a nut case by the media by not imposing a lockdown, it does not have any count (fraction of infections or deaths) as high as Italy. So, can Italy, or at least north Italy be already close to herd immunity?

In the week of the 2nd of April, they tested 773 people in Stockholm and 2.5% resulted infected. Extrapolating to the 9th of April, they estimate that $(7.5 \pm 2.5)\%$ of the population in the capital region were infected. Some detail can be found [here](#). Translating these numbers to the Italian data is not easy; it is not possible to use the number of infected people (the CFR) as the testing strategy in Sweden is different from Italy. The only possible common factor is the fraction of deaths over the whole population. There are two problems in using this fraction, one is the

different demographics of the two countries, the second is the fact that in Sweden all Covid deaths are considered while in Italy only those in hospitals are considered. The former difference can be corrected by comparing the fraction of over-70, the largest fraction of deaths being above this age. In Italy 17.2% of the population is over-70 while only 14.8% of Swedes are in the same age group. Therefore, we could expect a 16% ($17.2/14.8=1.16$) higher mortality in Italy simply due to demographics. A similar comparison done by experts can be found [here](#). It is difficult to set a correction for the different way of counting the deaths, the number of deaths in Italy should be higher than those reported but we could assume that in a non-overwhelmed region, all people with acute symptoms were transported in hospitals and therefore most deaths were counted.

Ok, moving to the math. Stockholm has a population of about 2.3 million and on the 9th had 486 deaths giving a fatality ratio (FR) of 0.02%. Please note that in this case the fatality is calculated over the whole population as this is a measure of the progress of the infection; this is different from the CFR where the denominator is the infected people. Actually, the two value are identical once the whole population is infected; at the current pace, Stockholm should have a CFR of 0.3, slightly higher than the lowest estimate reported in the review section but significantly lower than the Diamond Princes data suggested. The FR is then increased by 16% to be compared to the Italian regions fatality rates as described above. It is also crucial to say that Stockholm was not overwhelmed by the pandemic (a 400 places field hospital is still unused), so this number can only be compared to regions that are not overwhelmed, i.e. I will not show Lombardia. The results of this study are presented in Table 4. The extrapolation of the Swedish model is denoted as IP3 and both min and max values derived from the original statistical error are shown for completeness.

Table 4 Current fatality rate over the whole population and predicted fraction of infected based on the Swedish study

Region	Deaths / population [%]	IP3 min [%]	IP3 [%]	IP3 max [%]
Lombardia	0.1313	NA	NA	NA
Lazio	0.0044	0.9	1.4	1.9
Campania	0.0025	0.5	0.8	1.1
Sicilia	0.0029	0.6	0.9	1.2
Veneto	0.0200	4.2	6.3	8.4
Emilia-Romagna	0.0622	13.1	19.7	26.2
Piemonte	0.0350	7.4	11.1	14.7
Puglia	0.0073	1.5	2.3	3.1
Toscana	0.0101	2.1	3.2	4.2
Calabria	0.0026	0.6	0.8	1.1
Sardegna	0.0050	1.1	1.6	2.1
Liguria	0.0342	7.2	10.8	14.4
Marche	0.0280	5.9	8.9	11.8
Abruzzo	0.0028	0.6	0.9	1.2
Friuli Venezia Giulia	0.0172	3.6	5.4	7.3
Trentino-Alto Adige	0.0509	10.7	16.1	21.5

Umbria	0.0066	1.4	2.1	2.8
Basilicata	0.0037	0.8	1.2	1.6
Molise	0.0052	1.1	1.7	2.2
Valle d'Aosta	0.1090	23.0	34.5	46.0
Italy	0.0331	7.0	10.5	14.0
Italy without Lombardy	0.0171	3.6	5.4	7.2

The hope is that, if these numbers are confirmed, the northern regions are approaching infection rates that will allow a natural reduction of the spread of the virus and a return to a life closer to the pre-pandemic period. It must also be stressed that all regions in centre-south of Italy have a very low infection rate even in this model. So, a prudent and conservative approach is really needed to avoid a second wave in these regions. Italy really looks divided in two by these numbers and any policy should reflect them, to maximise the benefits in restoring personal freedoms in the north and to protect the fragile health system in the south.

Most affected provinces

While all the media was focussed on Bergamo due to the highest absolute counts of infected and deaths, the province that has the highest identified number of infected is Cremona followed by Lodi and Piacenza. To provide the estimates in these provinces I will need to make some additional assumption as some data is not available. For example, I will use the region demographic curve to calculate the number of 80-90 people affected in the province. The national male-female ratio is also used for the provinces. A difference in the infected demographic may be caused by different testing policies in different areas; overwhelmed areas may only test the severe cases while less affected areas may still test a larger spectrum of the population. Since older people are more affected, a bias in the tests based on the severity of the symptoms is also a bias in the demographic. Different provinces may also have different testing capabilities, resulting in a different sampling of the population. All these biases are likely limited by two factors, the regional based sanity system is likely to have provided similar resources per capita and that most cases are from the recent days when the whole region was under similar stress. The three models are presented in Table 5.

The conservative approaches described in IP1 and IP2 show that these provinces should already have a non-negligible fraction of the population infected, probably enough to already be having an impact on the transmission rate (as observed in Stockholm).

Since these provinces may have been overwhelmed, I used the national average (w/out Lombardy) and multiplied it by the ratio of IP1 between the province and Italy (w/out Lombardy) to scale the number of deaths for IP3. The results are shown in Table 5. According to this estimate, $(49 \pm 16)\%$ of people have been infected in the province of Cremona. This is very close to the levels required for herd immunity but there are many assumptions to reach these values which would need further scrutiny and more granular data to be sure they are correct.

Table 5 Estimates for the most affected provinces

Province	Population	Population M 80-90	Infected	Infected 80-90	Infected M 80-90	IP1 [%]	IP2 [%]	Fatality [%]	IP3 min [%]	IP3 [%]	IP3 max [%]
Cremona	358955	8434	5276	980	484	5.74	11.5	0.16	32.8	49.1	65.5
Lodi	230198	4526	2620	487	240	5.31	10.6	0.14	30.3	45.5	60.6
Piacenza	287152	7674	3106	512	253	3.29	6.6	0.09	18.8	28.2	37.6

Conclusion

I presented a way to analyse the available data on a regional base which better describe the fragmented Italian health system. Using the assumption that could be easily be verified by ISS which has more data, I provided 3 estimates of the number of infected as of the 16th of April. If more data could be made available, more precise prediction could be provided. Two predictions are rather conservative so can probably be considered lower limits. A recent study in Sweden provided a new way to estimate the infected that, if confirmed, would put the most affected provinces in Italy close to the levels required for herd immunity, or at least high enough to provide a significant rate reduction once the lockdown measures will be lifted. The Swedish are conducting a new study with higher statistics that will provide more accurate data and each Italian region should follow the same example as these tests are much cheaper than the mass tests that many regions are planning.

Let us hope that the last figures are correct and that we will be able to return to normal life soon.

Michele Faucci Giannelli

PS: Thanks to all people that provided feedback during the preparation of this post.