The Genie, Sampling and Covid-19

In an attic smelling of dust, mothballs and damp creeping slowly through the pages of old books, you find a huge copper urn with a very dusty lid and an old chamber pot perched on its edge. Through the clouds of dust you see that it is full to the brim with old copper coins, and a few gold sovereigns gleaming amongst them. Before you can reach in, the dust cloud settles into the form of a Genie, who tells you there are 10 million coins in the pot, mixed at random. You can have all the sovereigns if you can say whether there are more than 100,000 of them, or not. Get it wrong and you will have to polish all the coppers to a bright shine before you can leave the attic.You have half an hour.[1]

Now, the Genie told you that the coins were randomly mixed up in the urn. Suppose you use the chamber pot to shovel out a few thousand coins. If one percent of the coins in the urn were sovereigns, then about one percent of the coins in the chamber pot would also be sovereigns. Indeed whatever percentage of coins in the urn are sovereigns, roughly the same percentage of coins in the chamber pot will be.

So accepting that 'roughly the same' carries some risk of being badly wrong, you count the coppers and the sovereigns in the chamber pot, and work out the percentage that were sovereigns. Then work out how many sovereigns that percentage implies for an urn of 10 million coins... The chamber pot contained 2000 coins, and 40 were sovereigns – 2%. 2% of 10 million is 200 thousand. You answer the Genie, and the sovereigns are yours!

Obviously, by basing your answer on a sample, there was some chance of being wrong about the whole urn. But in this case some basic probability lets us work out the chance of getting 40 or more sovereigns in a sample of 2000, if the real percentage of sovereigns had been 1% or lower – it's less than 0.01%. So you were on pretty safe ground.

What if the Genie had not told you that the coins were randomly mixed? Without the 'randomly mixed' part, it could be that all the sovereigns were near the top, or near the bottom, or in a clump in the middle. Then you would have no reason to expect that the sample in the chamber pot would have roughly the same proportion of sovereigns as the whole urn.

That's because it is no longer the case that an individual coin has the same chance of ending up in the chamber pot irrespective of whether it is a sovereign or a copper. If we could sample the coins randomly, to ensure that a coin's chance of being in the sample was unrelated to its metal type, then we would have restored representativeness. For the urn example, that would be hard to do other than by physically mixing the coins – perhaps with a big stick. Let's move on.

Suppose that your Genie is altogether more malign, and you are tasked with estimating not the proportion of sovereigns in an urn of coins, but the proportion of people with Covid-19 in the population of the UK. There are 67 Million people in the UK and testing for Covid takes a lot longer than counting a coin. Testing everyone is not a sane option. Obviously the Covid cases do not occur randomly mixed among those 67 Million people. Infection occurs in clusters and varies geographically and by age group at any one time. So just picking a sample of people arbitrarily won't give us a sample in which we have any reason to believe that the percentage of infections is roughly that for the whole country.

As with the coins in the urn, if infections are not randomly mixed among people, we need to randomly sample people. We need each person to have an equal chance of being in the sample,

1 Slightly surprisingly some statisticians have written, in national newspapers, statements equivalent to asserting that the only way to rise to the Genie's challenge is to count all the coins. This would require the rapid improvisation of some very impressive coin counting machinery.

independent of whether they are infected or not. For people, that can be done without a big stick. Randomly pick people from NHS records, the electoral register, school rolls etc. Once you have this representative random sample, you test the people in it. The proportion of people infected in the sample will be roughly the proportion infected in the whole country. Moreover, because you sampled randomly, some basic calculations allow accurate calculation of how certain you are about your estimate (how different might it be if you estimated again with a different random sample, for example).

This random sampling approach is what the ONS survey does in order to estimate the Covid-19 infection rate. It is a very reliable well understood approach where the assumptions are clear.

What about Covid case numbers? These do not come from random population samples. We are in the situation in which the coins are not randomly mixed in the urn and we have not tried to mix them. So our sample proportions can not represent those in the urn. In addition case data is largely based on people tested because we think they might have Covid, perhaps through contact tracing or perhaps because they have cold symptoms. So it's as if, before counting, we immediately discarded from our sample any coins that at first glance looked like a copper. There is simply no way of estimating the proportion of people infected from these data alone. It's like trying to estimate the proportion of sovereigns, without bothering to assess the number of coppers.

Of course case data are *related* to the number of infections, but they are equally related to the ever changing way that they are collected and to many unmeasured factors. There is no basis for assuming that cases are proportional to infections in the general population - that is, for assuming that there is some constant number by which you could multiply cases to get a good estimate of infections. If you try to estimate infections from case data you will need to make a great many assumptions that are very hard to check. Estimates based on proper random sampling, as carried out by the ONS, are much better justified.

Basing decisions on changes in case data, as if these changes reflected change in infection rates, is therefore very difficult to justify. Cold like symptoms likely to lead to a test have a distinct seasonal pattern, for example. Similarly children, often asymptomatic for Covid, are much more likely to be tested on return to school as seasonal colds start to circulate. Track and trace is similarly evolving over time.

The bottom line: data are only as informative as the circumstances of their collection allow. If you'd rather be rich than spend an eternity polishing coppers, trust random sampling. If you want to make sane decisions on Covid-19, trust the ONS results in preference to changing case rates.