Statisitcs and Covid-19

Owen Ward

5/2/2020

Introduction

Covid raises many interesting questions related to data and has put statistics and data science in the publice eye more than any other event.

Will discuss a small selection of these:

- Difficulties of accurate testing
- Designing vaccines

The accuracy of testing

- Recently several unpublished articles have tried to estimate the prevalence of Covid-19 in the population using antibody tests.
- These have received widespread coverage in the media, https://www.cnbc.com/2020/04/17/santa-clara-covid-19-antibody-study-suggests-broad-asymptomatic-spread.html
- The above study estimated 50 times more people actually were positive than based on confirmed cases.
- These tests have accuracy difficulties which were not initially considered when the paper was released (and made the news).

Issues with disease testing

If you have a test for a disease, two things can go wrong:

- The test says you have the disease when you don't.
- The disease says you don't have the disease when you do.

Both of these cause issues, in different ways. We can get around some of these using Bayesian statistics.

Bayes Rule

We will use Bayes rule, which you may have seen before

$$P(D|P) = \frac{P(D \cap P)}{P(P)}$$

where D is the event that you have the disease, and P is the event that you test positive. We will use $\neg D$ to indicate the negative, that you cannot have the disease.

Both of these are binary events.

Conditional Probability

Suppose you have a test which tells you whether you have a disease or not. Given that you test positive, what is the probability you actually have the disease?

To do this, we need to know

- The sensitivity of the test
- The specificity of the test
- The true underlying proportion of the population who have the disease

Bayes Rule

We have

$$P(D|P) = \frac{P(D \cap P)}{P(P)}$$

where we can reuse this formula with A and B swapped to get

$$P(D \cap P) = P(P|D)P(D).$$

Similarly, by the law of total probability, we can re-write

$$P(P) = P(P|D)P(D) + P(P|\neg D)P(\neg D)$$

So we need to know $P(P|D), P(P|\neg D)$ and $P(D)$, which gives us $P(\neg D)$.

- $P(P|D)$ is the sensitivity of the test.
- $1 - P(P|\neg D)$ is the specificity of the test.

Sensitivity and Specificity

- Sensitivity is the probability the test is positive for the disease if you truly do have it.
- Obviously, we want this to be as high as possible.
- Specificity is the probability the test correctly gives you a negative result when you don't have the disease.
- We also want this to be high. But it maybe isn't as important as the sensitivity.

An example

- Suppose we have a test for some non serious disease, which 1 in 20 people have (so $P(D) = 0.05$)
- Let the sensitivity be 90%. So if you have the disease and get tested 10 times, you expect to be positive 9 times.
- Let the specificity be 80%. So, if you don't have the disease and you get tested 10 times, you expect that 2 of the times you will have a positive test.
- Suppose you test positive. What is the probability you have the disease?

An example

We have

$$P(D|P) = \frac{P(P|D)P(D)}{P(P|D)P(D) + P(P|\neg D)P(\neg D)}$$

where

- $P(P|D) = 0.9$
- $P(P|\neg D) = 0.2$
- $P(D) = 0.05$.

This gives

$$P(D|P) = \frac{(0.9)(0.05)}{(0.9)(0.05) + (0.2)(0.95)} \approx 0.19.$$

So if you test positive, you maybe shouldn't worry too much.

An example

- If we test 1000 people then we expect 50 of them to have the disease. We will correctly get positive tests for ≈ 45 of them.
- Of the 950 who do not have the disease we expect 2 in 10 false positives, giving 190 positives.
- In total, we would expect 235 positive tests, but only 45 of them will have the disease.
- Is this good? Is this good enough?

Increasing the specificity

If we increase the specificity to 0.95 then we can repeat these calculations and we get

- $P(D|P) \approx 0.49$
- If we test 1000 people now we will only expect to get 48 false positives.

What this means for Covid testing

- In the Santa Clara study done at Stanford, they test 3330 people.
- They get 50 positives, so raw $P(D) \approx 0.015$.
- They estimate the prevalence, after reweighting, is ≈ 0.03, with confidence intervals from 1.11% to 1.97%.
- They estimate the sensitivity is between 84% and 97%.
- They estimate the specificity is between 90% and 100%.

Covid Antibody Testing

Here, when we are looking at a rare disease, the specificity is particularly important. We saw in the example that most positives were false positives, unless the specificity is high.

- If the specificity is 98.5% and there are no true Covid 19 cases in the data, you would expect 50 false positives.
- It is difficult to know the true specificity, so any uncertainty in its value makes it plausible that the true disease prevalence is actually 0.
- This is just a very quick introduction to these ideas, see the references for some more detail.

Designing Vaccines

- Another extremely important topic with a pandemic like this is designing a vaccine for it.
- This is difficult and requires a type of statistics known as Causal Inference.
- To develop a vaccine we need to be able to prove its efficacy.
- Suppose you have a headache and you take an aspirin.
- If your headache goes away, is it because of the aspirin?
- Would it have gone away if you hadn't taken the aspirin?

Very basic Causal Inference

- To really answer this, you have to know the counterfactual. What would have happened had you not taken the aspirin?
- We never observe this. But we can get a good estimate of it by taking a properly random sample of people who all have headaches, giving half of them aspirin and half of them nothing. This is the basis of a clinical trial.
- The randomisation is very important here. We want the two random groups to be as similar as possible.

Simpsons Paradox

Suppose we have two treatments for kidney stones and we get the following data of the number of successes. We want to try determine which treatment is better.

	Treatment A	Treatment B
Small Stones	81/87 (93%)	234/270 (87%)
Large Stones	192/263 (73%)	55/80 (69%)
Overall	273/350 (78%)	289/350 (83%)

- Treatment A is better for both small and large stones but Treament B is better overall. How does that make sense?
- Treatment B was chosen by the doctors' for less severe cases. This skews the randomisation.
- This randomisation becomes even more challenging if there are multiple drugs being used, as is often the case now.

The Salk Polio trial

- Polio was a common disease among children in the first half of the 20th century, which can cause paralysis.
- Spread in waves within communities, leading to the closure of certain public areas from summer to summer.
- Vaccine proposed by Jonas Salk, decided to perform a large scale field trial.
- Polio was relatively rare (approx 1 in 2000) and so large number of people needed to confirm the vaccine was effective.
- Initial plan was to give the vaccine to every second grade student in the country, compare to students in first and third grade.
- Then doctors know who has got the vaccine, may influence how they diagnose/treat.

The Salk Polio trial

- In the end, half were given the above treatment while half were given a placebo method, where all were given a drug.
- Only the researchers knew who actually got the vaccine, while the rest got a placebo which did nothing.
- This placebo version is called double blind. People treating the children and administering the vaccine don't know what they are giving out.
- Which such large groups, randomly giving each person a treatment will lead to well balanced groups in terms of age, race, socio-economic status, etc.
- Trial contained nearly 2 million children, one of the largest clinical trials at the time.

References/Further Reading

- https://www.vox.com/2020/5/1/21240123/coronavirus-quest-diagnostics-antibody-test-covid
- https://statmodeling.stat.columbia.edu/2020/04/19/fatal-flaws-in-stanford-study-of-coronavirus-prevalence/
- https://statmodeling.stat.columbia.edu/2020/05/01/simple-bayesian-analysis-inference-of-coronavirus-infection-rate-from-the-stanford-study-in-santa-clara-county/
- https://www.medicine.mcgill.ca/epidemiology/hanley/c622/salk_trial.pdf