MODELLING RISK: “UNCERTAINTY IS ITS OWN RISK FACTOR”

White Paper by Alex Rodrigues for Philosophy Circle

Thank you to Henry Su for editing early drafts of this white paper

Often when evaluating a risk where one factor has high uncertainty people go astray using the simple heuristic “pick the most likely value”. If the most likely value is less dangerous than other possible values this can result in falsely underestimating the risk. This happens because uncertainty is often the largest source of risk in a poorly understood model.

Example of risk subject to uncertainty: COVID

To use a relevant recent instance of this type of analysis, let’s consider the risk of death for a person aged 60-69 with mild symptoms who just tested positive for COVID-19. According to the CDC at the time of writing there were 179,007 positive tests and 11,947 deaths in that age group[1]. In addition to lab confirmed cases a significant number of cases go undiagnosed which should also be considered. However, there is a lot of uncertainty about the exact number and distribution of these undetected cases.

At the time of writing a number of studies have done serological surveys of the population that suggest that the ratio of actual cases to positive tests is around 10:1. If we use this as the “most likely” value for the number of undetected cases we can calculate an overall mortality risk for a COVID positive person in the demographic group:

$$11{,}947 / (10*179{,}007) = \mathbf{0.67\%}$$

However this number gives a misleading sense of confidence - our understanding of undiagnosed cases is incomplete and there are a variety of ways the 10:1 ratio might not apply, for example:

- Maybe most of the undiagnosed cases are in young healthy people and actually most of the people in this older demographic group who contract COVID do eventually get tested
- Maybe the preliminary studies are inaccurate because they didn’t sample a representative portion of the population
- Maybe most of the undiagnosed cases are asymptomatic and the group of people sick enough to seek a test are a relatively separate group

The reality is the information we have is uncertain and the ratio of similar COVID cases to positive tests in this demographic group could be anywhere in a distribution. The ideal way to include uncertain assumptions in risk analysis is to evaluate each possibility based on it’s likelihood and danger level independently and then take a weighted average.

For the purposes of illustration, let's reduce this down to two possibilities, and give them a hypothetical level of likelihood:

A) There is a 60% chance that the ratio of cases to positive tests in this demographic group matches the preliminary studies and is 10:1
B) There is a 40% chance that the ratio of cases to positive tests in this demographic group is much lower for one or more of the possible reasons given above - it is actually 2:1

If we evaluate each possibility separately and then weight them by their likelihood we can estimate the risk accounting for this uncertainty:

60% * 11,947 / (10*179,007) + 40% * 11,947 / (2*179,007) = **1.74%**

By considering the two possibilities separately we arrive at a mortality risk that is almost 3x higher than when we used the most likely value alone. This is not a small difference and it is driven by the fact that with so much uncertainty a lot of the risk comes from the possibility our underlying assumptions were inaccurate. That is the first takeaway, for modelling risk:

"Uncertainty is it's own risk factor"

Averaging Risk vs Averaging Inputs

I want to draw your attention to the fact that using an average for the uncertain factor on its own is not the same thing as taking an average of the risk level. If we used the average value for the uncertain factor, we would still underestimate the risk, albeit not as badly:

60% * 10 cases/positive test + 40% * 2 cases/positive test = 6.8 cases/positive test (average)

11,947 / (6.8*179,007) = **0.98%**

Using the average value instead of the most likely value improves the estimate a bit - by incorporating more of the information - however it isn't the same as fully factoring in the uncertainty. This is because we haven't properly accounted for which assumptions are most dangerous.

In this situation a lower cases to positive tests ratio represents a more dangerous possibility. A lower number contributes relatively little to changing the average, however exactly how low it is affects the level of danger a lot. For example, having a 1:1 ratio in option B instead of a 2:1 ratio would almost double the combined risk level, but only moves the average down from 6.8 to 6.4. This highlights the importance of taking a weighted average of the overall risk and not the individual uncertain factors.

Takeaways

If we zoom out from the specifics of this statistical technique there is a more general lesson here:

> *“Be suspicious of overly confident models and explicitly consider the possibility that some assumptions of the model are wrong as its own risk factor.”*

If somebody says something is 99.9% unlikely based on reviewing 100,000 cases that’s probably about right. In contrast, if somebody says something is 99.9% unlikely based on partial information and some fancy modeling, the odds of the model itself making bad assumptions is often higher than 0.1%.

References

1. Coronavirus Disease 2019 Case Surveillance, United States, January 22–May 30, 2020 (Accessed July 19, 2020), Table 3., Centers for Disease Control and Prevention, https://www.cdc.gov/mmwr/volumes/69/wr/mm6924e2.htm