Hazard Analysis and Robustness

Hazard Analysis and Robustness

By James Camacho

Hazard Analysis

1. We have

$$k(51\%) - k(50\%) = \log_{10}\left(\frac{50}{49}\right) \approx 9 \cdot 10^{-3}$$

whereas

$$k(99\%) - k(98\%) = \log_{10}(2) \approx 0.3,$$

over 30 times larger of an improvement. If all risk measurements were in this format, it would be harder to multiply the reliability of multiple systems together. E.g. if two independent systems each have 1 nine of reliability, together they would only have about 0.72 nines of reliability, but to get that you have to convert back to regular probabilities.

2. If the risk is already known to be extremely large, and will increase incredibly while quantifying it more precisely, waiting to take action is still known to be worse than taking further action. For example, if the stock market starts crashing, and you know it will continue crashing, you don't wait to figure out exactly how much money you will lose--you take your money out as soon as possible to minimize loss.

3. 20% of the top 20% of the top 20% of land owners own 80% of 80% of 80% of all land, or 51.2% of land. So the top 1% of land owners own a little more than 50% of all land.

4. Taking logs gives $\log y = -\log 4 - 3x$ so the slope is -3.

5. (a) Known unknown (b) unknown known (c) known known (d) unknown unknown

6. Well it happened, so it's got to be more than one in 10^{50}. However, a Gaussian (thin-tailed) assumption would give a probability on the order of 10^{-89}, so it clearly is not a reasonable assumption. Long-tailed assumptions seem more reasonable for Black Mondays, and probably Friday the Thirteenth's as well.

Batman's IQ is about 6σ and Luthor's would be more than 8σ. The probabilities are on the orders of 10^{-10} and 10^{-17} respectively. There's about 10^{10} humans alive, so you would expect about one person to be as smart as Batman, but Luthor's intelligence is entirely unrealistic.

7. (a) boxing helmet = preventative, ice pack = protective
(b) lifeboats = protective, compartmentalized hull = preventative
(c) eating healthy = preventative, chemotherapy = protective
(d) teaching honesty = preventative, catching lies = protective

8. It's exactly 16x more costly to enact protective measures than preventative. Okay maybe not exactly, but it's cheaper to prevent disasters than try to clean up after them.

9. (a) positive feedback
(b) self-organization
(c) micro-macro dynamics
(d) butterfly effect
(e) positive feedback
(f) positive feedback
(g) positive feedback

10. Political, social, and economic pressures.

11. Rapid COVID tests are safe but not reliable. Missiles are reliable but not safe.

12. I disagree with this. Benevolent actors will cooperate and try to rectify their mistakes, leading to a negative feedback cycle of damage. Maleveolent actors will try to exploit mistakes in the system, and with every exploit are able to do more damage, leading to a positive feeback of damage.

Robustness

1. 'Tis $d\varepsilon$, $\varepsilon\sqrt{d}$, and ε respectively. The perturbation is larger for $p = 1$.
2. Beta(1, 1) is uniform, Beta(5, 5) has a peak at the center, and Beta(0.5, 0.5) has a peak at the corners.
3. (a) False (b) True (c) True (d) True
4. The best ℓ_2 norm will be minimized when x is the projection of x_0 onto the line $w^T x + w_0 = 0$, and $x = x_0 + \delta w$. I.e.

$$w^T(x_0 + \delta w) + w_0 = 0 \iff \delta = \frac{-w_0 - w^T x_0}{\|w\|_2^2}.$$

Plugging this in gives $x_{\text{adv}} = x_0 - \left(\frac{w^T x_0 + w_0}{\|w\|_2}\right)\frac{w}{\|w\|_2}$, or (c).

5. (a) False
   (b) False
   (c) True
   (d) True
6. (a) False, but it's almost good enough as models usually have similar weaknesses even if they have completely different parameters.
   (b) True
   (c) True
   (d) False
7. (a) False, but only because they norm the error after every step of PGD. That's basically the only difference.
   (b) True. E.g. cropping a bunch or cutting out scenes.
   (c) We have

   $$\nabla_x L = -yw \frac{e^{-yw^T x}}{1 + e^{-yw^T x}}.$$

   The fraction is always positive, so the sign is $-y\text{sign}(w)$.
   (d) True
   (e) True
8. (a) False
   (b) False
   (c) False
   (d) False
   (e) True