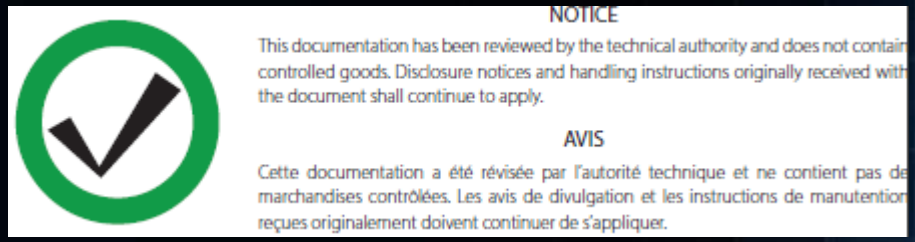
DEFENCE R&D DÉFENSE

Preventing Premature Conclusions

Analysis of Human-In-the-Loop Air Combat Simulations

30 August 2012
Matthew MacLeod
Centre for Operational Research and Analysis



Outline

- The issue
- Randomness and expectation
  - Issues with small samples
  - Missile outcomes
  - Issues with cognitive bias
  - Displaying uncertainty
- Streaks
  - Example
  - Impact
- Implications to practice
  - Why not get rid of the randomness?
  - Conclusions
- Recommendations
  - to the analyst
  - to the client (trial director, tactician, requirements developer)

The issue

- Realistic, human-in-the-loop simulation of few-on-few fighter combat has become not only feasible, but the preferred (or even only) option for comparing current and future tactics and aircraft
- Stochastic elements can easily be introduced for ‘realism’ or to avoid exploring every possible outcome – but human intuition with regards to average results can be problematic
  - As for intuition combined with a room full of alpha personalities…
- Data is generally closely protected, complicating analysis
  - A simpler analysis presented *in situ* is often better than a much delayed follow-up analysis that cannot be easily distributed
- What is the analyst’s role in this scenario?
- How do you convey uncertainty to trial participants, given often misleading intuition?

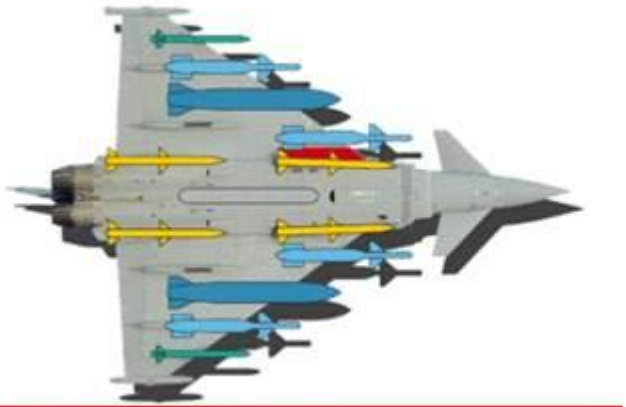
Randomness and expectation: context

- A completely deterministic simulation is not practical for a realistic encounter
  - Probabilistic elements are necessary to represent uncontrolled factors (e.g. weather) and unpredictable factors (e.g. relative aspect)
- In particular, despite improvement of missile kinematic models, some end-game factors must still be treated via a stochastic P_{kill}
  - i.e. the simulation will determine whether the missile reaches intercept, but in the end does a 'dice roll' (pseudo random number generator) to determine whether the target is then killed
- Most clients' intuition is that by fixing the parameter of this binomial variable, they will be able to fairly compare aircraft/weapons/tactics across runs
  - Often some lip service is paid that comparisons may not be 'statistically valid, but…'

The issue: small samples

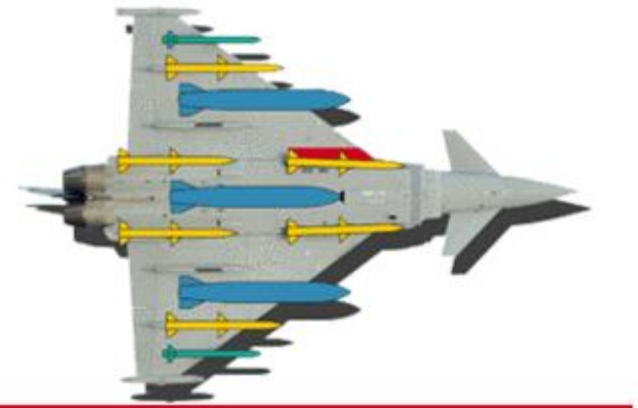
- Modern fighter combat (whether virtual or real) defined by few v. few encounters
  - Not unlike many sports, even if you know the strategies, the players, and the training, one can only hope to know the *expected* (that is, average) outcome of an encounter
  - We're hopefully not planning on conducting wars of attrition with our small numbers of expensive aircraft
  - What meaning does exchange ratio have in vital point protection?
- Moves towards multi-role weapon loads and internal carriage reduce number of air-to-air missiles available
  - No matter how much better each missile is, there will always be some chance of failure – and fewer trials to average over
  - Variance in outcomes of a few missiles carried by a few aircraft tends to swamp the effect of the variables you're trying to compare

BVR load-out examples



- 4 x Laser guided bombs
- 4 x BVRAAM
- 2 x SRAAM
- 27mm Mauser
- 2 x 1000 litre fuel tank
- Laser designator pod

Air Interdiction mission configuration



- 6 x BVRAAM
- 2 x SRAAM
- 27mm Mauser
- 3 x 1000 litre fuel tanks

Air Superiority mission configuration

- Assuming that most encounters are to take place beyond visual range (BVR)
  - Short range encounters are a contingency
- A typical multi-role fighter may have four or six BVR missiles loaded
- A typical formation size may be four or six fighters (or even two)
- Sixteen or twenty-four missiles per formation per encounter not a particularly large sample to average over

Expectations: missile outcomes

P_{kill}	E[Kills] ± 2 σ	
	24 missiles	16 missiles
10%	2.4 (+2.9/-2.4)	1.6 (+2.4/-1.6)
25%	6.0 ± 4.2	4.0 ± 3.5
50%	12 ± 4.9	8.0 ± 4.0
75%	18 ± 4.2	12 ± 3.5
90%	22 (+2.0/-2.9)	14 (+2.0/-2.4)

- P_{kill} estimates for actual missiles are highly sensitive – showing a wide range
- Note spreads as wide as 7 to 17 kills for P_{kill} of 50% and 24 shots
- Even the narrowest spreads are four kills wide
- What is a ‘reasonable’ number of kills to plan for in a trial, or in reality?

Issues with cognitive biases – part 1

- Even knowing the numbers, it can be hard to fight intuition
- Both laypersons and trained scientists have been shown to believe in the ‘law of small numbers’ – that small samples should be close to the average, just like large samples
- Even more counter-intuitive is the phenomenon of ‘regression to the mean’
  - In any trial with repeated random components, a highly successful result is very *likely* to be followed by a less successful result, and vice versa
  - This is not due to the universe ‘averaging out,’ but is simply due to more of the probability distribution being on one side of the previous result
  - This is problematic when comparing two different things in subsequent runs – it is hard to shake the qualitative impression that the second run went much better or worse than the first, even if the difference is due to the random outcomes

"The reliance on heuristics and the prevalence of biases are not restricted to laymen. Experienced researchers are also prone to the same biases—when they think intuitively."

Amos Tversky and Daniel Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, pp. 1124–1131, 1974.

Options for displaying uncertainty

- Given the need to fight our natural tendencies, it is important to be able to display (and re-display) uncertainty
- In some senses, which is not shown is more important than what is shown
  - If numbers (e.g. kill ratio, missiles per target) are flashed up for different runs, people are naturally going to fixate on them – but their meaning may be suspect
  - Just because something is easy to count, doesn't mean it's important
- Can explore both tabular and graphical representations

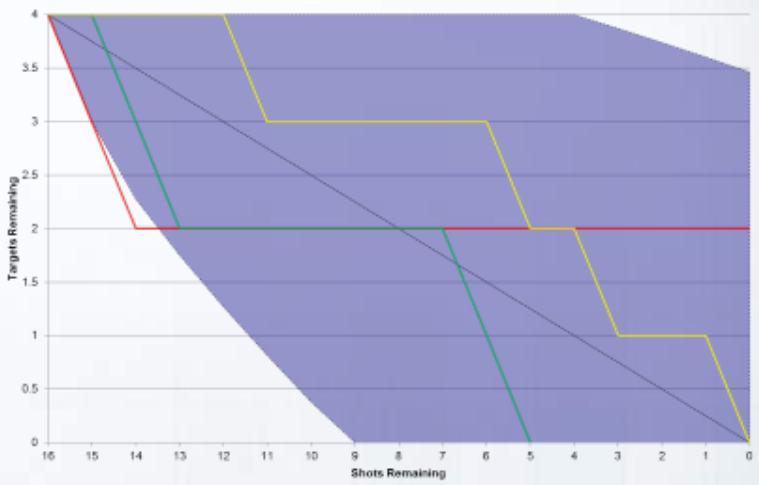
Display option 1 – big ugly table

P_{kill} = 50% P(Success) = 1 - P(Targets avoiding more than (Missiles - Targets) shots)

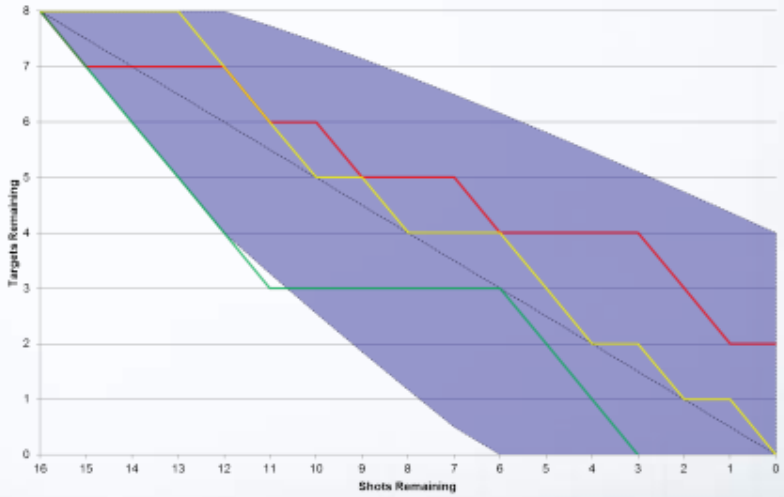
	Targets							
Missiles	8	7	6	5	4	3	2	1
16	59.82%	77.28%	89.49%	96.16%	98.94%	99.79%	99.97%	100.00%
15	50.00%	69.64%	84.91%	94.08%	98.24%	99.63%	99.95%	100.00%
14	39.53%	60.47%	78.80%	91.02%	97.13%	99.35%	99.91%	99.99%
13	29.05%	50.00%	70.95%	86.66%	95.39%	98.88%	99.83%	99.99%
12	19.38%	38.72%	61.28%	80.62%	92.70%	98.07%	99.68%	99.98%
11	11.33%	27.44%	50.00%	72.56%	88.67%	96.73%	99.41%	99.95%
10	5.47%	17.19%	37.70%	62.30%	82.81%	94.53%	98.93%	99.90%
9	1.95%	8.98%	25.39%	50.00%	74.61%	91.02%	98.05%	99.80%
8	0.39%	3.52%	14.45%	36.33%	63.67%	85.55%	96.48%	99.61%
7	0.00%	0.78%	6.25%	22.66%	50.00%	77.34%	93.75%	99.22%
6	0.00%	0.00%	1.56%	10.94%	34.38%	65.63%	89.06%	98.44%
5	0.00%	0.00%	0.00%	3.13%	18.75%	50.00%	81.25%	96.88%
4	0.00%	0.00%	0.00%	0.00%	6.25%	31.25%	68.75%	93.75%
3	0.00%	0.00%	0.00%	0.00%	0.00%	12.50%	50.00%	87.50%
2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	25.00%	75.00%
1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	50.00%

Display option 2 – probability regions

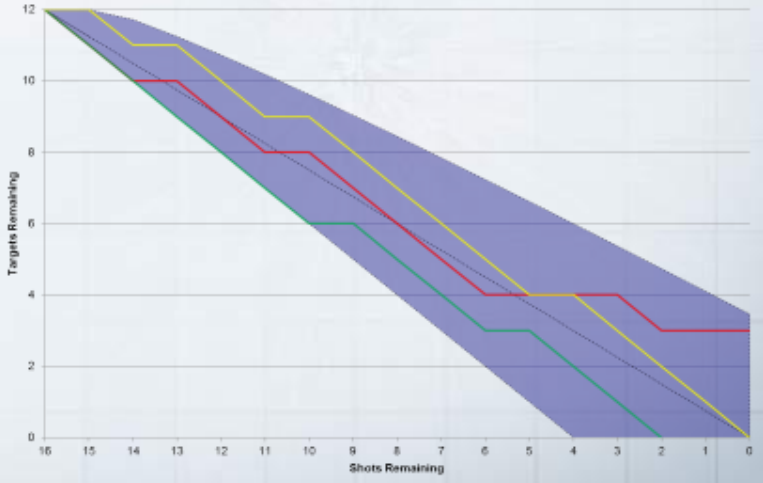
- *x* – shots remaining
  y – targets remaining
- Diagonal line tracks average number of kills
- Shaded area is 2 std dev in *height* – note *y* axis is scaled for P_{kill}
  - Height at 0 shots remaining is the same for 25% and 75% - see next slide



P_{kill} = 25%

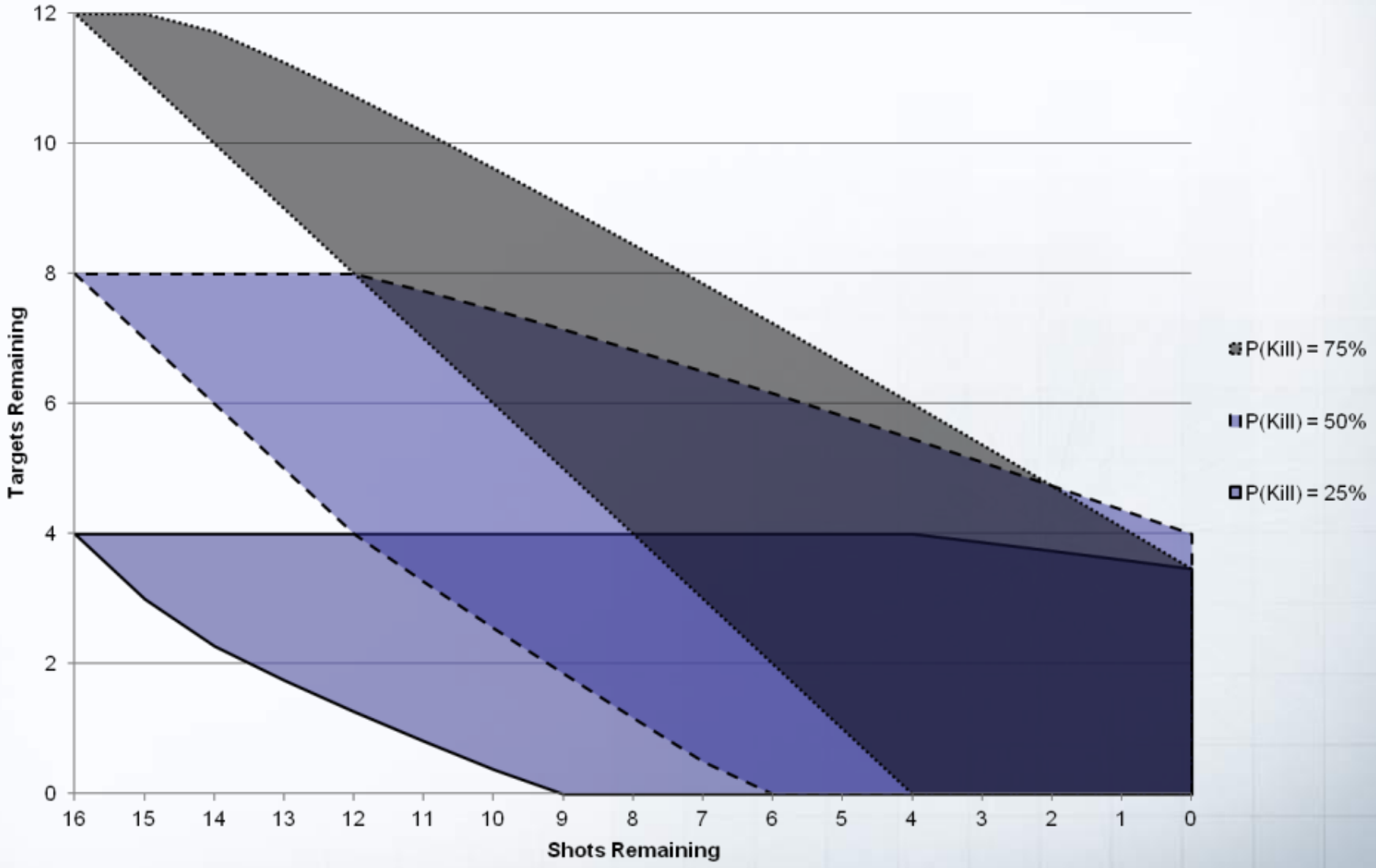


P_{kill} = 50%



P_{kill} = 75%

P_{Kill} example comparison



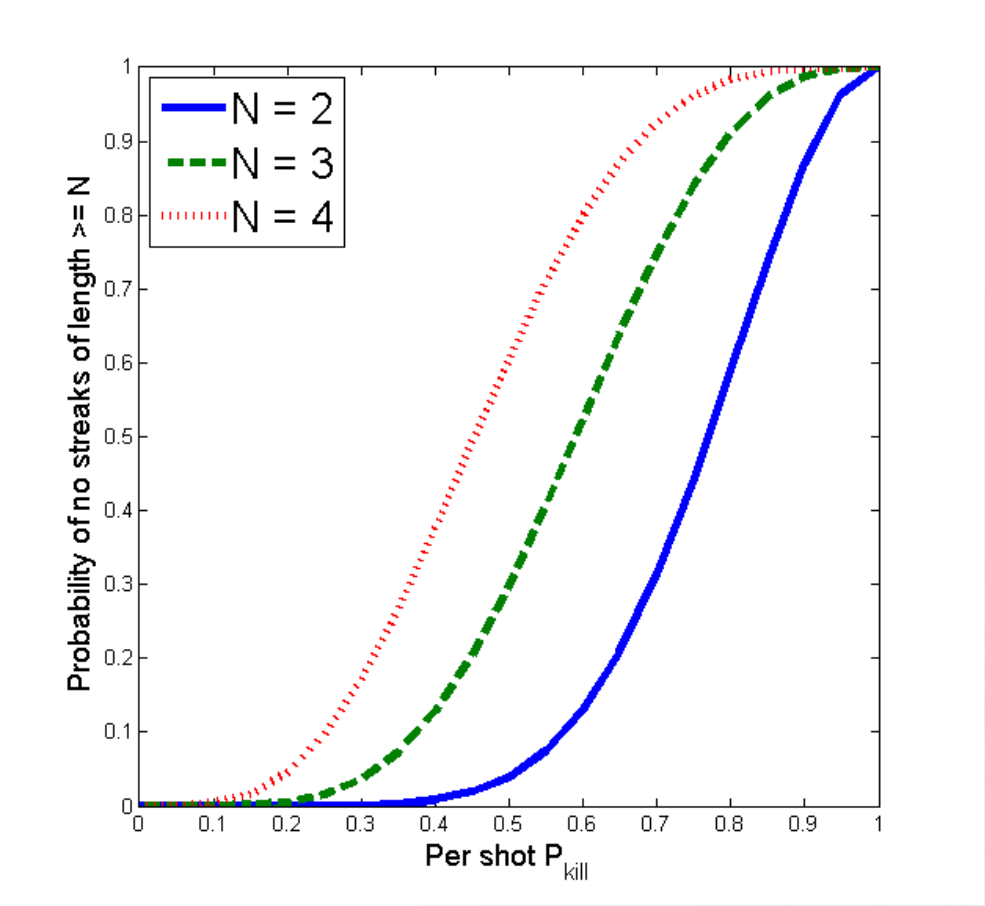
Some notes on expectations

- If the P_{kill} is *only* representing the end-game factors, consider also the likelihood of intercept
- If success means a Blue:Red kill ratio of 0:*N*, the question starts to look pass/fail for a given scenario
  - Can the fighter/weapon/tactic handle an enemy force of size *N*?
  - But given what we've just seen, even if *only probabilistic missile outcomes are considered*, there will be some quantifiable risk of failure

Streaks

- Humans have been shown to consistently mistake the likelihood of streaks in random processes
- The “gamblers fallacy” refers to underestimation of streaks
  - Simplest example is a fair coin
  - Easy to determine the probability that the next flip will be the same is 0.5
  - When asked to choose a ‘random’ looking sequence, it has been shown that we choose those with a 0.7-0.8 chance of alternating between flips
  - Conclusion is that we assume sequences will ‘even out’ substantially more quickly than probability tells us
- The “hot hand” refers to overestimation
  - Common in sports, where we tend to believe that a player who is performing well is more likely to continue, and vice versa
  - Distinction is we believe the person has agency, whereas a coin does not
- Missile firings are vulnerable to both interpretations

Example of streak likelihood



- *x* – per shot P_{kill}
- *y* – probability of *not* having a miss streak of length *N* in 16 shots

Impact of streaks

- If tactic assumes roughly average performance per volley/wave, may be more vulnerable than expected
- Don’t forget that there are non-probabilistic reasons for missile failure as well
- If participants try to write-off an ‘unlucky’ streak, important to be able to quickly tell them exactly how probable it is
  - Easily calculated as $(1-P_{kill})^n$
  - Can also emphasize that in repeated encounters, likelihood of at least one of them having a streak goes up quickly

So what do we do from here?

Why not just get rid of the randomness?

- Participants may become bored and lose focus
- Even worse, may adjust their behaviour and ‘fight the simulation’
  - e.g. may not feel the need to do realistic battle damage assessment

- Even more pressing, however, is that the uncertainty and streakiness are not an artifact of the simulation – real encounters will have the same small sample sizes
- Developing the air crew’s understanding of probability not just essential for the simulation results, but to their actual application of air-to-air tactics in training and combat
- The trick is to fight the ‘that will never happen’ response

Conclusions

- The statistician who loves neat and tidy comparisons may be disappointed – the point is (in)decision support
- Importance of qualitative lessons – you can still recognize Wayne ~~Gretzky~~ Rooney, even if their team doesn't always win
  - But those qualitative lessons need to be tempered by identifying actual outliers in the underlying process
  - Watch out for regression to the mean
- As the number of fighters and missiles in encounters get smaller, need to focus more on risk than on averages

Recommendations to the analyst

- Avoid presenting averages, or at least unannotated averages
  - Focus instead on noting whether missile outcomes for a particular run were within one or two standard deviations
- Highlight runs that were particularly ‘lucky’ or ‘unlucky’
  - Important to temper enthusiasm for using lucky runs as models of tactical success
  - Can actually be worse when a tactic is thrown out based on an unlucky run, especially when participants have prejudged the outcome
- Pick your battles – the trial is rarely designed for your own benefit
  - Focus on rectifying the most egregious issues, and let small things go
- Be aware of your own biases
  - The analyst is just as susceptible as the layperson to well-known cognitive biases and forming opinions based thereon
  - Being self-deprecating and using analogies to less charged subjects will also help to ensure one does not appear too preachy

Recommendations to the client

- Match-ups between aircraft should be looked at in terms of probability of success and risk levels, rather than exchange ratios
  - Assumption is not that we're fighting a war of attrition, but are interested in protecting a valuable asset
  - Plan tactics based on number of firing opportunities, use probabilities as an overlay to that
- The robustness of one's tactic to a streak of 'unlucky' missile results will continue to be important
  - Streaks will happen even for high P_{kill}
  - Acquiring missile with better 'dice' may help, but not the whole solution

Questions?


DEFENCE
R&D
DÉFENSE