Calling Experiments Using Decision Theory

Zach Flynn

Significance levels

- 5% significance level is arbitrary.
- But it's useful to have a standard, something fixed.
- Without a standard, tend to squint and ship.
  - "Directionally positive"
  - "We're going to 'test and roll'"
  - "Positive (not stat sig)" ... <proceeds to act exactly the same as if it were stat sig>
- Too strict of a standard encourages this and is bad.
  - If I never reach the standard, I'm not just going to not ship anything for a year.
  - We have this intuition that there's some reason to be biased towards status quo, but not a lot in biz.
- Well-trod stuff, but what does it mean tactically? How do we quantify it? If not 5%, then what?

Statistical Decision Theory

- We’re not primarily trying to measure stuff.
- We want to make decisions.
- Decision theory is thinking about how to map datasets to decisions.
- In general, have to model “ambiguity” as well as “risk”
    - Risk = preference between risky and certain outcomes, but with known probabilities of outcomes (do you like 50-50 coin flip of $100 vs $0 or $45 certain?)
    - Ambiguity = preferences when we don’t know the probabilities of outcomes (I’m going to roll this die and it might be loaded but we don’t know)
- We don’t know how likely various outcomes are.

Expected Loss/Utility

- A neat thing about making statistical decisions using the Expected Loss setup is it has a nice axiomatization as far as risk:
  - Completeness. For any lotteries, A and B, either A >= B or B >= A.
  - Transitivity. A > B, B > C => A > C.
  - Continuity. A > B > C => there exists *p* such that: pA + (1-p) C < B < (1-p)A + p C.
  - Independence (key). For any B and p < 1, A <= C iff (1-p) A + p B <= (1-p) C + p B.
- Bayesian world handles ambiguity with a prior.
- We're going to try to use the Bayesian Expected Loss setup to come up with decision rules in Frequentist Inference.

An Expected Loss Framework for Hypothesis Testing

- Going to cast frequentist inference in an expected loss setup.
- Idea is to make decision *d* about the effect of an experiment in two levels:
  - Have we learned anything from the data? (*h*)
    - If we haven't (*h=0*), then we go with some default belief (d = s).
    - If we have (*h=1*), then we pick some decision (d) about the effect.
- The true treatment effect is *t*.
- Loss: $L = (1 - h)(t - s)^2 + w\ h\ (t - d)^2$, w measures how much weight we put on making a call vs sticking with the default
- Expected Loss: $E[L \mid X] = (1 - h)\ E[(t - s)^2 \mid X] + wh\ E[(t - d)^2 \mid X]$
- Idea is to model what we actually do: decide the results can be "trusted" and then just pretend the point estimate is the truth.

Expected Loss

- h = 0
  - $E[L(h=0) \mid X] = E[(t - s)^2 \mid X]$
  - $Var[t \mid X] = Var[t - s \mid X] = E[(t-s)^2 \mid X] - E[(t-s) \mid X]^2$
  - $E[L(h=0) \mid X] = E[(t - s)^2 \mid X] = Var[t|X] + E[(t - s) \mid X]^2$
- h = 1
  - $E[L(h=1) \mid X] = w\, E[(t - d)^2 \mid X]$
  - Because d only shows up here, minimum loss is at $d = E[t|X]$.
  - So… $E[L(h=1)|X] = w\, Var[t \mid X]$

Deciding Whether To Make A Call

- $h = 1 > h = 0$ when:
  - $w\,\mathrm{Var}[t \mid X] > E[(t - s)^2 \mid X] = \mathrm{Var}[t|X] + E[(t - s) \mid X]^2$
  - $(w - 1)\,\mathrm{Var}[t \mid X] > (E[t|X] - s)^2$
  - $(w - 1) > (E[t|X] - s)^2 / \mathrm{Var}[t \mid X]$
- The right hand side looks like a Wald test statistic.
- Now, in Expected Loss world, we have in mind expectations over beliefs, so $E[t|X]$ and $\mathrm{Var}[t|X]$ are most closely something like a posterior...
- But these are actually kind of close to the frequentist estimates under a broad class of priors via the Bernstein-von Mises theorem.

Intuition

- So, let's just interpret the right hand side as a Wald test.
- The significance level we want solves:
  - Let F be the cdf of a $X^2(1)$ random variable. 1 - F(w-1) = effective significance level.
  - So, what is the standard 5% level? w - 1 = 3.84 => w = 4.84 — almost five times more weight on loss from making the wrong decision relative to just ignoring the data.
  - Reducing the weight to say w = 3, puts the significance level around 0.16.
  - w = 1 => just ignore statistical error and always go with the point estimate.
    - The deeper naivete here: ignoring SE and just going with the point estimate is optimal under some alternative criteria, see Hirano and Porter (2009).
- The idea is to use this framework for thinking about reasonable significance levels because it's easier to think in terms of relative square loss.

Why?

- Want to balance caring about noise with the need to, at some point, make a call at scale.
- Want to set a uniform standard.
- The method doesn't really take a stand on how much we want to bias towards our "default"—but I will!
    - Only making a decision by mistake 5% of the time "seems" way too low. We now have a way to say that it's basically put 5x weight on mistakes we make by making a decision.
    - We want to have w > 1 because we want some status quo bias, I think. First, do no harm and all that.
    - If the expected loss is 3x smaller going with the estimate than *s*, then I'm cool with it.

Are we actually increasing error?

- This will kind of depend on the company, but…
  - Power is not really that moveable most places. You can't really wait two months to make a decision and even if you could, you might not have the power (at 5%) you need to really be confident about rejecting for effect sizes you'd like to detect.
  - Adjusting the MDE to match the maximum possible sample size can be helpful for setting expectations with stakeholders, but it's not really going to do anything once the test is live.
  - Sure, we could be like "don't run underpowered tests", but then what will we do? Not run tests and just ship stuff and, in the best case, do some kind of causal inference on the rollout. These come with their own problems.
  - So, suppose we run these underpowered tests and, naturally, they usually don't show us anything significant. De facto, we start behaving suspiciously like w = 1 folks.
- So, we have to ask: if we use a 0.16 significance level, and we can then properly power the tests are we making more errors relative to… whatever we would be doing otherwise?
- By putting less weight on the error from making a call, maybe paradoxically, people pay more attention to standard errors. If you're always insignificant, they kind of blend into the background. If the good experiments are usually significant, then lack of significance is more surprising, evidence we aren't doing what we'd like to do => iterate.

Wrapping Up

- We know 5% is arbitrary — and we can afford to take more risks.
  - Of course, “it depends.” Boeing, etc.
- So, concretely, how can we justify other levels?
- An expected loss model gets us to some reasonable significance levels.

References

- https://faculty.washington.edu/kenrice/testingrev2a.pdf