What Makes a Good Hypothesis

Alvin Wan

November 2021

Your project presentation should be oriented around the following four categories: **EDA, Hypothesis, Experimental Design, Interpretation**. The underpinning of your project, across all 4 steps actually, is a good hypothesis: get this right, and the design for the rest of your project will flow more naturally. We expect your hypotheses to have these two traits:

1. **The hypothesis can be proven or rejected.** This criteria sounds obvious in retrospect but is easy to forget or miss. Make sure that your hypothesis can be tested, considering the data and time you have.

    - One common mistake is requiring a non-existent dataset or feature. For example, your hypothesis may require household income for every address in San Francisco. Given this data is not public, the hypothesis would be un-testable. However, the fix is to simply consider less granular data – in other words, use household income per *neighborhood*, which *is* available online: link.
    - Or, your hypothesis may require an attribute too vague to have data for. For example, your hypothesis may require knowing how "well connected" each person is. This definition is not easily quantifiable and may be subjective. However, the number of LinkedIn connections is easily quantifiable and is possibly correlated with your idea of well-connected. Note that you can simply say "For the purposes of this project, we define connectedness as the number of LinkedIn connections", leaving more refined or expanded definitions theoretically to "future work". *Caveat: Note it is possible to solidify such an abstract definition in a full research paper, where you've considered many interpretations and draw upon large bodies of previous work. It's tough to do in the time constraints you're given.*

2. **The hypothesis has a plausible negation.** In other words, take the reverse of your hypothesis, and ensure that this reversed hypothesis is plausibly true. If the reversed hypothesis is obviously false, then the *original* hypothesis is uninteresting. This helps distinguish an "uninteresting" from an "interesting" hypothesis. Here are examples to help you calibrate what "obviously false" means; there's some nuance:

- Avoid the obvious. The hypothesis "COVID death rates are positively correlated with COVID cases" is uninteresting, since its negation "COVID death rates are *not* correlated with COVID cases" is obviously false. A patient dies from COVID only after contracting COVID, by definition, so the two statistics *must* be correlated. This makes the original hypothesis uninteresting.
- Know your context. If you change the scenario, such a basic correlation test could become valuable: Let's say you were trying to identify if a variant of COVID is a "variant of concern". Then, this check may be useful: "Candidate COVID variant is positively correlated with death rates": If the variant is not correlated, it may not be affecting mortality rate and isn't immediately deadly; if it is correlated, the variant should be labeled and should raise red flags for public health officials if found.
- Explain your background knowledge. The hypothesis "COVID death rates are negatively correlated with number of hospital visits" is actually a decent hypothesis. Its negation "COVID death rates are positively correlated with number of hospital visits" may seem obvious at first, but background knowledge of the situation tells us otherwise: Given hospitals were packed with the most serious of COVID patients, it's quite possible that hospital visits increased other patients' exposure to more deadly strains of COVID. The question would be whether this increased exposure is enough to counteract the quality of medical care. *The most interesting hypotheses are those previously seemed obviously false or obviously true and were then proven otherwise. However, this is tough to pull off. We don't suggest pursuing this, but technically, if you find evidence in your EDA, you could pursue such a hypothesis. For the purposes of grading, make sure to present this evidence.*

The two above metrics for a good hypothesis are more nebulous and commonly missed. Here are some other criteria that you may have already considered:

- The hypothesis has support. The support may be a prior belief that you validate through EDA.
- The hypothesis relates two variables. Ideally, the two variables are not simply columns in the data we provided but ones you compute from the columns we provide, or ones from external sources.

The above are the minimum working requirements for a hypothesis. However, there are other nice-to-haves that help spice up a research project. These ideas have their own fair share of risks that take a lot of consideration to avoid though, as attractive as they sound on the surface: Showing the obvious is not obvious. Leveraging a clever use of data. Showing a relationship where none was thought to exist before.