CS 294-115 Algorithmic Human-Robot Interaction **Fall 2016**

Lecture 8: Experimental Design Continued

Scribes: Smitha Milli and Jeremy Warner

In this lecture we discussed what makes for a good experiment.

8.1 Good experiments are controlled.

Controlled experiment: An experiment in which an experimenter assigns experimental units to treatments, as opposed to a simply observational study.

Why are controlled experiments good? Because they allow us to avoid the natural confounds that tend to arise in observational studies such as the examples below.

Example 1: Estrogen supplements. In one observational study the use of estrogen among post-menopause women was found to be highly positively correlated to health. There was a confound here though - the superior health of the subjects who took estrogen could have merely been because they were more health conscious to begin with. When a controlled experiment that randomly assigned estrogen intake to women was conducted it was shown that taking supplemental estrogen actually has a severe negative effect on health outcomes.

Example 2: Left-handed lifespan. Another observational study found that people who are left handed tend to die nine years before their right handed counterparts. However, this was actually because the proportion of left-handed people has increased over time, due to relaxed cultural pressure around being right-handed.

8.2 Good experiments avoid confounds.

Confound: a variable whose effect cannot be distinguished from the effect of the independent variable.

Examples: Health conscientiousness, age, gender, genetic factors, socioeconomic factors, technical expertise, test problem bias (e.g. testing motion planning in unrealistic environments with random blocks), and metaparameter optimization.

Tools to avoid confounds

1. Randomize.
   (a) Randomly assign experimental units to conditions. This really means completely random. For example, do not assign the first half of your participants to condition A and the second half to condition B because there could there could be a difference in the population of the first half versus the second half. Perhaps, people who come earlier tend to be students and people who come later tend to be working individuals getting off from a 9-5 job.
   (b) Blocking. Another way to account for confounds is to vary the independent variable(s) within *blocks* of experimental units that are homogeneous with respect to a confounding factor. For

example, you may create separate blocks for males and females and then test types of drugs within those two separate blocks to reduce variability due to gender.

2. Run a within-subjects experiment. You can avoid many confounds in a within-subjects trial because all subjects see all conditions, so you are not prone to confounds stemming from a difference in the populations of subject that see one condition versus another. However, in a within-subjects trial there is a need to take care to avoid ordering effects in the conditions shown to subjects.

   For example, if you are presenting two tasks, A and B, that are similar to each other to users, you shouldn't always do B after A because the user may have learned how to do B better through their experience in A. You can use the technique of *counter-balancing*, randomly picking an ordering of conditions. Unfortunately, counter-balancing scales requires $n!$ orderings for n conditions, so this can get infeasible when there are too many conditions.

3. Optimize the baseline. This is specifically addressed towards the common confound of metaparameter optimization, i.e., the fact that we often spend hardly any time optimizing baselines compared to the time spent optimizing our own algorithms. One good way to test for fairness here is to have someone else optimize both the baseline and your algorithm.

4. Stack the cards against yourself. This is a technique that addresses the common confound of test problem bias. Different algorithms have different environments or objectives they will perform comparatively better or worse on. When comparing your algorithms to an alternative algorithm, rather than using the objective your algorithm was designed for, you can show superior performance on the objective that the alternative algorithm was designed for. This gives a stronger reason to believe that the performance of your algorithm will generalize to other environments/objectives. For example, if you are comparing CHOMP and RRT, rather than only using path smoothness as a metric of comparison, you could also test path length.

8.3 Good experiments are reliable.

Good experiments are *reliable*, i.e., they have low variance. You can test and improve the reliability of your experiment through repeated measurements and multiple direct parallel measurements.

8.4 Good experiments have construct validity.

Good experiments have *construct validity*, i.e, they measure what you say that they measure. One tool for ensuring construct validity is to include both objective and subjective measures.

For example, in measuring how human-like a robots actions are: a subjective example could be to use the Likert scale and ask them how human like the robot tends to seem (ranking on a scale of 1-7). An objective measures could include asking participants what they think the robot will do next in the middle of its performance.

Good experiments have both high reliability and construct validity. If you only have high reliability, you may come to the wrong conclusion with a very high amount of certainty, perhaps due to including a confound in your experiment. This is shown in the lower right corner of Figure 8.1, along with other combinations of reliability and construct validity. In each case, the objective truth lies at the center of the large circle, and the smaller circles represent results from experiment trials.

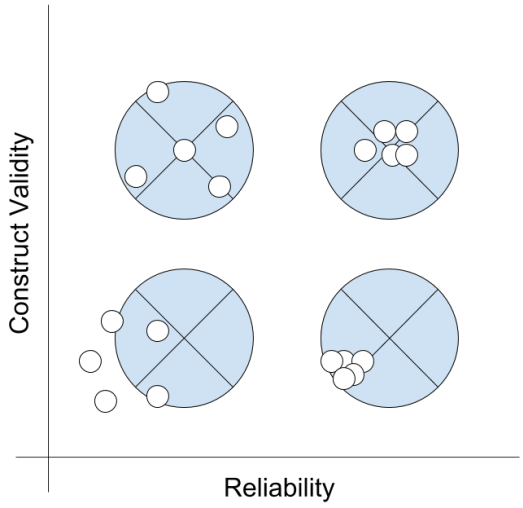

Figure 8.1: Reliability vs. Construct Validity

8.5 Good experiments have external validity.

Are our conclusions generalizable? In robotics, we often would like to show generalization around across problem instances, problem types, and algorithms.

A key step here is to extract what the independent variable is in your experiment and test that in a way that matches the generality of the conclusion you want to make. Anca brought up an example from some of her early work that sought to show that goal sets improve motion planning algorithms. However instead of testing on multiple motion planning algorithms, the tests reported were only on CHOMP, so the only conclusion that could be reached was that CHOMP with goal sets was better than CHOMP without goal sets. In order to make the claim that goal sets improve motion planning algorithms in general, a stronger test that could have been reported would be a comparison between TrajOpt with goal sets and TrajOpt without goal sets.