Lecture 5: Experimental Design

Scribe: *Patricia Chaffey*

9/7/18

1 Definitions

Theory: a system of ideas intended to explain something
Hypothesis: a supposition made on the basis of limited observations or intuition, intended as a starting point for further investigation

Examples used throughout the lecture:

- We want to test algorithm A against algorithm B
- We want to test drug A against drug B

2 The Components of a Study

1. Independent Variables
   - Algorithm A or Algorithm B
   - Drug A or Drug B
2. Dependent Variables
   - Performance metrics
   - Symptoms
3. Population
4. Hypothesis
   - Relationship between independent and dependent variables
   - Generating a hypothesis typically through testing an observation or intuition
   - Pilot studies are a good way to generate hypotheses
   - Note: having a clear hypothesis before one begins allows the researcher to create control groups

3 Running the Experiment

1. Between subjects

   - Large differences between subjects may lead to high amounts of potential variation
   - Sample size must be large enough

2. Within subjects

   - Ordering effect: A followed by B may influence users when trying out B, having already experienced A
   - In cases where the ordering effect is not very strong, randomizing the order of A and B across subjects may address the issue

Covariant: factors that vary in the population that we cannot control

- Example: in the group of students testing out the algorithms on the robots, some of the students are CS students and others are biology students, so their prior experience with robots will be varied

Confounds: covariates that affect what you measure

- Example: In a study about estrogen treatment, there was a significant correlation to positive health. A possible confound in the study might be that people who take estrogen treatments are generally more health conscious and opt to eat healthier, or work out more, and therefore are overall healthier

4 Biases

1. **Researcher Bias:** How an experiment is explained to a subject may inadvertently influence how they approach the task

   - Counteract this by using a script, or by having someone unfamiliar with the inner workings of the experiment run it

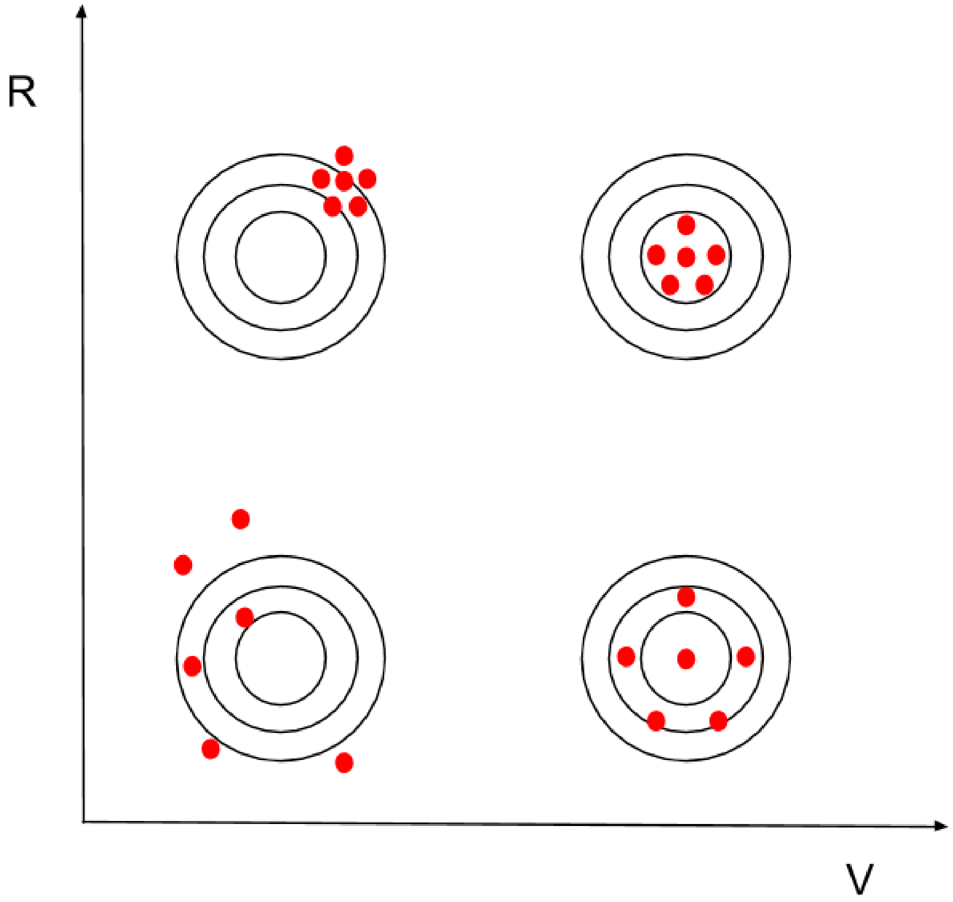
2. **Annotation Bias:** measurements taken inconsistently or otherwise influenced by some means

   - Counteract this by either having people unfamiliar with the experiment take the measurements, or by having machine measurements (e.g. a system that automatically detects when a human is moving or standing still)

5 Two Important Qualities Desired in Experiments

1. Validity: measure the right thing
2. Reliability: measure the thing right (as in, your results shouldn't wildly differ each time you run the experiment)

The below image demonstrates reliability vs. validity, where the red dots can be considered data points.



6 Likert Scale

A common way to see the Likert scale used in HRI is to present a statement, such as ”I find the robot trustworthy,” and then ask participants to rate their level of agreement with it. A scale may look like [1-strongly disagree] up to [7 - strongly agree]. However, this is not the best way to retrieve data. A better system uses multiple statements, which leads us to Muir's questionnaire.

1. To what extent can the robots behavior be predicted from moment to moment?
2. To what extent can you count on the robot to do its job?
3. What degree of faith do you have that the robot will be able to cope with similar situations in the future?
4. Overall, how much do you trust the robot?

Where the participants may then be asked to rate their level of agreement with each statement.

7 Study Execution

1. Consent form
2. Execute study (record everything in the event that you want to investigate an unanticipated effect or observation in the future)
3. Measure objective/subjective responses
4. Evaluate data

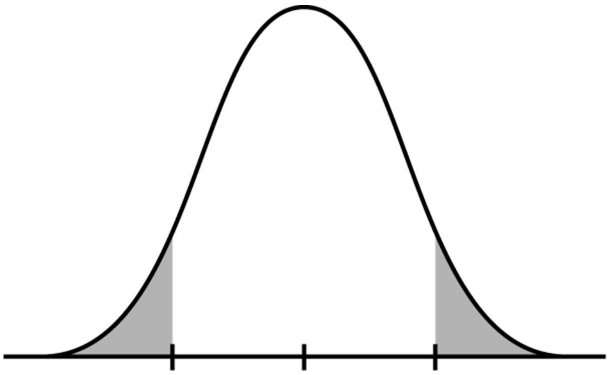
An aside on the Stanford Prison Experiment - The purpose of this experiment was to investigate the pyschological effects of perceived power, in the context of a prison housing prisoners and officers. The experiment was abandoned after six days, as participants (college students) left midway, and others embraced their roles as guards or prisoners. Early reports of the experiment claimed that subjects took to their roles, where those in the guard position enforced strict measures that led to psychological abuse of the subjects in the prisoner roles

8 Statistical Analysis

Given result, how do we check that they are reasonable?

- H_0 : the null hypothesis, which can be thought of as the default assumption for the model generating the data
- H_A : the alternative, which we accept as the best explanation if we reject the null hypothesis

The significance level, indicated by the shaded region on the graph, is the probability of rejecting the null hypothesis when true. In general, when running a hypothesis test, the null hypothesis can be rejected for the entire population when the sample statistic is different enough from the null hypothesis. 'Different enough' is defined by our own assumption that the null hypothesis is true, the significance level, and the data itself.



8.1 Type I Error

Type I Error: the rejection of a true null hypothesis, also known as a false positive.

P(1 or more errors in 20 comparisons) = 1 - P(no error) = 1 - 0.95^{20} = 0.64
where 0.95 is the probability of not making a Type I error.

An aside on the Bonferroni correction - when testing multiple hypotheses, the chances of a rare event occuring increases, which also increases the likelihood of incorrectly rejecting a null hypothesis. To counteract this, the Bonferroni correction tests each individual hypothesis at a significance level of α /m, where α is the desired overall alpha level and m is the number of hypotheses.

Theorem: X_i, ..., X_n independent normal random variables with means μ_1, ...,μ_n and variances σ_1^2,...,σ_n^2

$\gamma = \sum_{i=1}^{n} c_i X_i + f$
$\gamma \sim \mu[\sum_{i=1}^{n} c_i \mu_i + f, \sum_{i=1}^{n} \sigma_1^2, c_1^2]$
$E[X + f] = EX + f$
$\gamma[cX] = c^2 \gamma[X]$

Z-test
A Z-test should be used on test statistics that follow a normal distribution. For the following example, assume we have a random sample from observations with unknown mean and unknown variance.

Null hypothesis (H_0) : if we know the average time task situation, then

- H_0 : $\mu = \mu_0$
- H_1 : $\mu \neq \mu_0$

$\overline{X} = \sum_{i=1}^{n} x_1/n \sim N(\mu_0, \frac{\sigma^2}{n})$
$\overline{X} - \mu_0 \sim N(0, \frac{\sigma^2}{n})$
$Z = \frac{\overline{X}-\mu_0}{\frac{\sqrt{n}}{\sigma}} \sim N(0,1)$

The next quiz will be based on the papers assigned.