CS 294-115 Algorithmic Human-Robot Interaction **Fall 2015**

Lecture 6: Experimental design continued

Scribes: Amy Pavel and Kiwoo Shin

2 What makes a good experiment (cont.)

2.5 External validity

Good experiments have good external validity. **External validity** means that conclusions drawn from the experiment are generalizable. One way to achieve good external validity is to sample from the target population (e.g. don't test on only 3 year olds if you would like your conclusion to apply to all ages). Here are a few methods for sampling from the target population in the context of robotics:

- Apply the experiment to problems representative of the real world
- Try the experiment on a variety of problems
- Try your insight on algorithms as well as problems (e.g. in motion planning problem, try goal sets with more than one algorithm including CHOMP, GS-CHOMP, Trajopt, GS-Trajopt and so on)

2.6 Factorial experiments

Good experiments are factorial. A common pitfall in many experiments is to observe the dependent variable while changing several independent variables at once. In this case, we cannot isolate the effect of each independent variable. A fully **factorial experiment** consists of two or more factors each with discrete possible values, and whose experimental units take on all combinations of these levels across all such factors. For instance, if we have an experiment with two factors (e.g. goal sets, and representation) each with two levels (e.g. waypoint representation and RKHS representation) we would test all combinations:

	No goal sets (No GS)	Goal sets (GS)
waypoint	No GS, waypoint	GS, waypoint
RKHS	No GS, RKHS	GS, RKHS

Such factorial experimental design lets us isolate effects of different factors, or **main effects**, and detect **interaction effects**, which occur when the effect of one of the variables differs depending on the level of the other variable.

When conducting experiments in HRI, especially if they involve human participants, keep in mind the size of the fully factorial experiment when considering the number of factors you want to test and the desired precision.

3 Analyzing experiments

This section contains a brief overview of statistical analysis methods people expect to see in HRI papers. Use this section to figure out what type of statistical analysis is appropriate for your experiment. However, statistical testing is an active area of research, and this is not an exhaustive list. To learn more about how to report each type of test, refer to the APA's guidelines for reporting significance.[1]

In this section **IV** stands for independent variable and **DM** stands for dependent measure.

3.1 1 IV, 2 levels, 1 DM, within-subjects design

This example experiment includes one independent variable with two levels (goal set vs. no goal set), one continuous dependent measure (cost) and two experimental units (motion planning problems). We want to test if using goal sets results in significantly lower cost paths on motion planning problems.

MP Problems (Exp. Units)	Level (IV)	Cost (DM)
1	0 (No GS)	10
1	1 (GS)	8
2	0	11
2	1	7

Because we have a within-subjects design, we can also represent the experiment as follows:

MP Problems	Cost difference
1	2
2	4

To test if the cost difference between not using goal sets and using goal sets is greater than 0, we can use a **t-test**. A t-test considers the sample mean, $\overline{x}$, the sample standard deviation $\overline{s}$, and the sample size N.

$$t = \frac{\overline{x}}{\overline{s}/\sqrt{N}}$$

Using t and N, the sample size, we can find the p-value. A low p-value tells us it is unlikely that we would observe a test statistic t if the null hypothesis (i.e. the difference between using goal sets is not greater than 0) were true. We choose a p-value ahead of time (usually 0.05) and report significance only if our p-value is below the selected threshold. To determine the N we need to detect an effect with given degree of confidence, we can use **power analysis**.

If the dependent measure is categorical rather than continuous, we use a **Chi-squared test** (χ^2) instead of a t-test.

Note: There's difference between not being able to claim with any statistical confidence that two things are different and being able to claim that these two things are the same. To test if two things are the same, we can use an **equivalence test**. Equivalence tests are used in medicine to claim, for instance, that a newly invented drug has the same effect as the original.

[1] Please refer http://evc-cit.info/psych018/Reporting_Statistics.pdf for more about APA style of reporting statistics. This link includes how to report result of t-test and ANOVA.

3.2 1 IV, 2 levels, 1 DM, between-subjects design

If the t-test is between-subjects, we use the following equation to obtain t:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\overline{s_1}^2}{N} + \frac{\overline{s_2}^2}{N}}}$$

3.3 $>$ 1 IV or $>$ 2 levels, 1DM, between-subjects design

Suppose we had an experiment with 2 independent or more independent variables, and we created a factorial experiment like the one displayed in Table 2.6. We cannot use 6 t-tests to compare all conditions without adjusting the p-value because the **multiple comparison problem** implies that each subsequent test for significance increases our chance of finding significance in error. As an illustration, here we find the chance of making at least one error in 100 comparisons assuming a p-value of 0.05.

$$\begin{aligned} P(\geq 1 \text{ error in 100 comparsions}) &= 1 - P(\text{always correct}) && (6.1)\\ &= 1 - 0.95^{100} && (6.2)\\ &= 0.9941 && (6.3) \end{aligned}$$

We are very likely to make an error! One conservative method to adjust the p-value for such errors is the **Bonferoni correction**. Using the Bonferoni correction, we obtain a corrected p-value by dividing the original p-value by the number of comparisons.

However, we usually use an analysis of variance, or **ANOVA**, to detect main effects and interaction effects if we have more than one independent variable. If interaction effects are detected, we follow up with a post-hoc analysis (e.g. **Tukey**).

3.4 $>$ 1 IV or $>$ 2 levels, 1DM, within-subjects design

For within-subjects studies we use a **repeated measures ANOVA** to detect any overall differences between related means.

3.5 Multiple DMs

Much debate here. MANOVA is designed for this. But when DMs highly correlate you average them into a score and do ANOVA. When they are highly different then independent ANOVAs.

3.6 Tools for statistical testing and reporting

- **JMP**: Allows you to enter data with data types and run relevant tests, includes recommendations
- **Minitab**: Simple statistical testing software, only works on Windows
- **R**: Programming language and software environment for statistical computing and graphics
- **APA**: Guidelines for reporting significance