36-309/749
Experimental Design for Behavioral and Social Sciences

Sep. 15, 2015
Lecture 3: Experimental Design Principles

Review of One-Way ANOVA (§7.2)

- Ideal Model: Each member of a population has a potential ***quantitative outcome*** for each of k (≥2) different treatments. For each level of the ***categorical explanatory variable*** there is a different distribution of the outcomes. Each distribution is Normal in shape, has population means labeled μ_1 to μ_k, and has identical spread (σ^2). The errors (individual deviations from the population means) are independent of each other.
- H_0: $\mu_1=\ldots=\mu_k$. H_A: at least one population mean differs from the others (Not H_A: $\mu_1\neq\ldots\neq\mu_k$, because, e.g., $\mu_1\neq \mu_2= \mu_3=\ldots= \mu_k$ is not in either hypothesis.)
- Experiment: Randomly select N subjects from the population. Randomly assign treatments to the subjects. Today we focus on equal "n", so n=N/k. Avoid correlated errors.

One-way ANOVA, cont.

- A useful statistic: $F=MS_{between\_groups}/MS_{within\_groups}$.
- $MS_{within}=SS_{within}/df_{within}$ is an estimate of σ^2 ***whether or not*** the null hypothesis is true. $df_{within}=$ [ ].
- $MS_{between}=SS_{between}/df_{between}$ is another estimate of σ^2 *if* the null hypothesis is true, but ***larger*** otherwise.
  $df_{between}=$ [ ] .
- Under the null hypothesis, experimental repetitions will give F statistics that vary, but center around 1. The "null sampling distribution" of this F statistic, ***if the assumptions are true***, is the theoretical distribution called $F_{a,b}$ where a is the numerator df and b is the denominator df. (For those who want to be exact, E(F)=b/(b-2).)

One-way ANOVA, cont.

- For any given experiment, under ***each*** alternative hypothesis there is an ***alternative*** sampling distribution of F. These vary from slightly to the _____ of the null sampling distribution to far off to the _____.
- From the position of the one observed F statistic in its theoretical ***null*** sampling distribution, we can find the p-value (significance level) for that one experimental run.
  One design→ distrib. of F's → distrib. of p-values.
- If the p-value is _____ than our pre-chosen alpha (α), e.g. 0.05, then our results are "surprisingly uncommon" for similar experiments in which the null hypothesis is true. The decision is: reject the null hypothesis. For a randomized experiment, conclude that treatment ***causes*** a change in the mean population outcome.

One-way ANOVA, cont.

- Reject H_0: either correct or a "type-1" error
- Retain H_0: either correct or a "type-2" error
- The type-1 error rate is α. The type-2 error rate depends on the power of the experiment.
- Violation of assumptions → true sampling distribution of F changes → p-value calculated from the standard sampling distribution are wrong (also SEs and CIs). The degree of this problem depends on the **robustness** of the test.

One-way ANOVA, cont.

- With k=2, t-test or ANOVA works and the p-value is the same.
- Optional "proof" for $n_1=n_2=n$:

$$t^2 = \frac{(\overline{Y_1}-\overline{Y_2})^2}{{s_p}^2\left(\frac{2}{n}\right)} = \frac{(2\,d_B)^2\left(\frac{n}{2}\right)}{MS_W} = \frac{MS_B}{MS_W} = F$$

- SPSS Output (ANOVA table)

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	63.021	1	63.021	5.828	.020
Within Groups	497.458	46	10.814		
Total	560.479	47			

Some Principles of Research Design

- "The goal of any research design is to arrive at clear answers to questions of interest [about the populations] while expending a minimum of resources." –Ramsey and Shafer
- Identify sources of experimental variation, i.e., things that make the error variance (σ^2) larger, and consider controlling these. (§8.5)
  - Subject to subject variability
  - Measurement variability
  - _____ variability
  - _____ variability

Principle 1: Assure interpretability

- **Principle**: Avoid criticism about ***causality*** if the p-value turns out to be ≤0.05. (§8.1)
  - **Internal validity** indicates that we have good reason to believe that it is the differences in treatment that ***cause*** the differences in outcome. Assure no other IV is unbalanced across treatment groups, i.e., prevent confounding.
    - **Randomize <u>treatment application</u>**
    - "Expectation" can be a confounder! Use **blinding**, including a **placebo**, if possible to avoid the possibility that differences are due to expectations about treatment rather than treatment itself.
    - Differences in outcome could be biased by differential drop out.
  - To be sure what caused the effect, have the treatment groups differ in only one aspect, if possible and appropriate.
  - Use a **control group**, if possible, to have something to compare effects to.

Principle 2: Construct Validity (§8.2)

- Are the definitions of the DV and IVs well defined, reliable, and reproducible?
- Do the DV and IVs really measure what we want them to and what we call them?
- E.g., calling the sum of the number of parties you were invited to but did not attend a measure of "shyness" is debatable.

Principle 3: Promote Broad Inference

- **External validity** or **generalizability**: Prevent your experiment from having limited impact through criticism about what population your sample represents, particularly if the p-value turns out to be ≤0.05. (§8.3)
  - Can we generalize from one age group to others? race? gender? nationality? education?
  - Can we generalize from a carefully controlled environment to the real world?
  - Can we generalize from carefully controlled treatment application to the real world?
  - If possible, **randomize <u>subject selection</u>** (totally distinct from randomizing treatment assignment, above). Avoid convenience samples and other sources of a "sampling bias".

Principle 4: Promote Power

- **Principle:** Improve your ability to detect real differences. Make "not statistically significant" results meaningful. Avoid criticism if the p-value turns out to be >0.05. (§8.5)
  - **Control** the four sources of **variation**.
  - Measure whatever pre-treatment characteristics that you can't control, and appropriately include those measurements in your model as **factors** or **covariates**.
  - **Blocking**: Group similar subjects into "blocks" and randomize treatment application within those blocks. Analyze in a way that "pools" results across blocks. Examples of blocks include grouping by experience, apparatus, location, etc. Blocks are an added factor whose significance we don't bother to test.

Power Principle, cont.

- Use **within-subjects designs** where each subject is his or her own control, so that the subject-to-subject variation is mathematically isolated, reducing the "effective" error. (But this design may not be possible, or may introduce other problems.)
- Assure that your treatments are strong enough (compared to control).
- Assure that you have enough subjects.
- Note: It may be worth "trading off" some generalizability for more power.

Principle 5: Do the right test

- Check EDA before running a test and residuals afterwards to assure that the mode assumptions of the test are met (considering robustness to assumption violation); otherwise the p-value, SE and CI lose their meanings.
- Solutions: transformation, weighting, better means model, alternate (more robust) procedures (often less powerful). (§8.4)
- Also, avoid uncorrected **multiple testing** (§13.3).

Example 1

This field experiment tested the effect of a monetary incentive on speeding behavior. Using GPS technology integrated with GIS referenced speed limit information, eight vehicles were instrumented in a manner that allowed real time knowledge of vehicle speed relative to the speed limit. Fifty participants drove these vehicles, with each individual driving his or her assigned vehicle for a four week trial. During one week, 40 participants experienced an automated feedback system, which provided visual and auditory alerts when they sped five or more mph over the limit. Twenty of these 40 individuals experienced a monetary incentive system during their second and third weeks of driving. Ten participants were in a control group that experienced neither system. The percent of time speeding is the DV, the treatment group is the main IV, and speed limit is a blocking variable.

(A field experiment to test the effects of automated feedback and monetary incentive on speeding behavior, Ian J. Reagan, Old Dominion University, 2011)

Example 2

Eleven children with early focal brain lesions were compared with 70 age-matched controls to assess their performance in repeating non-words, in learning new words, and in immediate serial recall, a triad of abilities that are believed to share a dependence on serial ordering mechanisms. The children with brain injury showed substantial impairment relative to controls in the experimental tasks, in contrast with relatively unimpaired performance on measures of vocabulary and non-verbal intelligence. These results support previous reports that there are persistent processing impairments following early brain injury, despite developmental plasticity. They also suggest that word learning, non-word repetition, and immediate serial recall may be relatively demanding tasks, and that their relationship is a fundamental aspect of the cognitive system.

(Phonological memory and vocabulary learning in children with focal lesions Gupta, et al., Brain and Language, 87:241 , 2003)