Applied Statistical Methods II

Introduction to the Design of Experimental and Observational Studies

Looking forward:

- A very broad overview of issues in design that we will cover in more detail for the rest of the semester.
  - Experimental vs. Observational
  - Basic concepts of experimental studies.
  - Most popular experimental studies.
  - Some topics in observational studies.

Experimental vs. Observational Studies

- Experimental Studies:
  - We have control over the assignment of treatments.
  - Randomization can be used to limit the effect of potential confounders.
  - Well designed experimental studies can be statistically easy to analyze.
- Observational Studies:
  - We have no (or little) control over the assignment of treatments.
  - Confounding variables almost always effect observed difference between treatment groups.
  - Need more complicated statistical machinery to determine treatment effects.

Experimental Study Example

- Study to determine the effect of baking temperature on the volume of a bread from a packaged mix.
  - Experimental factor is temperature.
  - Factor is the design equivalent for predictor and independent variables in linear models.
- We are interested in four different temperatures: low, medium, high, and very high.
  - There are four factor levels for temperature.
  - We have four treatment groups. Unique pattern of factors.
- There are 20 packages of mix which we can use in our experiment.
  - Objects to which treatments are assigned are called **experimental units**.
  - **Randomization** is essential in assigning treatments to experimental units.
  - Potential confounders, such as age on shelf, should be independent of treatment assignment.

Observational Study Example

- A business school wants to know the effectiveness of a teaching seminar.
- All of the 110 faculty are invited, but not required to attend.
    - 63 attend, 47 do not.
    - Independent variable of interest is attendance.
- At the end of the semester, students rate each professor's performance.
    - On a scale from 1-7 with 7 being optimal.
- Naive approach is to compare the mean score among attendees vs. non-attendees.
- If goal is to understand effectiveness of teaching seminar, what could go wrong here?
    - Would have to control for potential confounders.
    - Maybe the non-attendees dislike teaching a priori.

Observational Study Example

- A business school wants to know the effectiveness of a teaching seminar.
- All of the 110 faculty are invited, but not required to attend.
  - 63 attend, 47 do not.
  - Independent variable of interest is attendance.
- At the end of the semester, students rate each professor's performance.
  - On a scale from 1-7 with 7 being optimal.
- Naive approach is to compare the mean score among attendees vs. non-attendees.
- If goal is to understand effectiveness of teaching seminar, what could go wrong here?
  - Would have to control for potential confounders.
  - Maybe the non-attendees dislike teaching a priori.

Mixed Studies

- Studies can include both experimental and observational factors.
- A study to compare performance of two different training programs.
  - Outcome is employee performance.
  - The company has 3 different plants.
  - Randomly assign training programs to employees.
- Are you awake: What are independent variable(s)? Dependent variable? Experimental units?
  - Experimental units are the employees.
  - Training program is an experimental factor.
  - Plant where the employee works can be an observational factor.
- What if you randomly assign training programs to plants instead of employees? What are experimental units? Potential confounder(s)?
- Certain randomization designs can be used to eliminate biases from the observational factor.

Mixed Studies

- Studies can include both experimental and observational factors.
- A study to compare performance of two different training programs.
  - Outcome is employee performance.
  - The company has 3 different plants.
  - Randomly assign training programs to employees.
- Are you awake: What are independent variable(s)? Dependent variable? Experimental units?
  - Experimental units are the employees.
  - Training program is an experimental factor.
  - Plant where the employee works can be an observational factor.
- What if you randomly assign training programs to plants instead of employees? What are experimental units? Potential confounder(s)?
- Certain randomization designs can be used to eliminate biases from the observational factor.

Mixed Studies

- Studies can include both experimental and observational factors.
- A study to compare performance of two different training programs.
  - Outcome is employee performance.
  - The company has 3 different plants.
  - Randomly assign training programs to employees.
- Are you awake: What are independent variable(s)? Dependent variable? Experimental units?
  - Experimental units are the employees.
  - Training program is an experimental factor.
  - Plant where the employee works can be an observational factor.
- What if you randomly assign training programs to plants instead of employees? What are experimental units? Potential confounder(s)?
- Certain randomization designs can be used to eliminate biases from the observational factor.

Mixed Studies

- Studies can include both experimental and observational factors.
- A study to compare performance of two different training programs.
  - Outcome is employee performance.
  - The company has 3 different plants.
  - Randomly assign training programs to employees.
- Are you awake: What are independent variable(s)? Dependent variable? Experimental units?
  - Experimental units are the employees.
  - Training program is an experimental factor.
  - Plant where the employee works can be an observational factor.
- What if you randomly assign training programs to plants instead of employees? What are experimental units? Potential confounder(s)?
- Certain randomization designs can be used to eliminate biases from the observational factor.

Single vs Multi -factor studies

- First two examples are single factor studies.
- Last example is a two-factor study.
  - For now we will only concentrate on multi-factor experimental studies.
- Two-factor crossed study:
  - Want to know how temperature and concentration affect a chemical reaction.
  - Temperature has three factor levels: low, medium, high.
  - Concentration has two factor levels: low, high.
  - Want to explore different concentrations under different temperatures.
  - Have 2x3=6 treatment combinations.
- Usually addressed with an ANOVA model with interaction.
- Let's write out full model. Potential issues?

Nested Factors

- Want to know the effect of a human operator on productivity.
- We have three different plants.
- We select three operators from each plant.
- Two-factor study:
  - Plant has three factor levels.
  - Operator has nine factor levels (i.e. 9 people).
  - Each operator works at only 1 plant.
- What is the problem with a standard linear model here?
- Operator is said to be nested within plant.
  - Each operator only operates in one plant.
  - Have 9 different treatment conditions.
  - How do we want to compare these conditions?
- Will extend the ANOVA models to include random effects.

Nested Factors

- Want to know the effect of a human operator on productivity.
- We have three different plants.
- We select three operators from each plant.
- Two-factor study:
  - Plant has three factor levels.
  - Operator has nine factor levels (i.e. 9 people).
  - Each operator works at only 1 plant.
- What is the problem with a standard linear model here?
- Operator is said to be nested within plant.
  - Each operator only operates in one plant.
  - Have 9 different treatment conditions.
  - How do we want to compare these conditions?
- Will extend the ANOVA models to include random effects.

Power and Sample Size

- Assume you conducted a single-factor study with four treatment groups.
- Your goal is to determine if there are any differences among the four groups.
- Fundamental question: how many subjects do you need to achieve a certain power? Let's investigate...
  - Recall that power is the probability of detecting any differences given that the amount of difference between the groups is Δ.
- The power will depend on:
  - The effect size Δ.
  - The error variance σ^2.
  - The number of subjects n.
  - Number of replicates per treatment group.
  - From simple example: it appears a balanced design leads to estimates with minimal variance.
- When asking someone for money to do an experiment, these are important issues that must be addressed.

Power and Sample Size

- Assume you conducted a single-factor study with four treatment groups.
- Your goal is to determine if there are any differences among the four groups.
- Fundamental question: how many subjects do you need to achieve a certain power? Let's investigate...
  - Recall that power is the probability of detecting any differences given that the amount of difference between the groups is Δ.
- The power will depend on:
  - The effect size Δ.
  - The error variance σ^2.
  - The number of subjects n.
  - Number of replicates per treatment group.
  - From simple example: it appears a balanced design leads to estimates with minimal variance.
- When asking someone for money to do an experiment, these are important issues that must be addressed.

Randomization for Assigning Treatments

- Consider the bread example.
- We have n=20 experimental units (mixes).
- We have 4 treatments: low, medium, high, & very high temperature.
- Want to have five replicates per treatment.
- Should we just pull the packages off the shelf and give the first five low heat, next five medium, . . . ?

Simple Randomization

- Randomly generate a number for each observation.
- Sort these numbers from lowest to highest.
- Assign low temperature to the lowest 5, medium to the next lowest 5, ...
- Can be done with a random number generator.
- SAS has PROC PLAN that can be used for different randomization.

```
data randomize;
do bag=1 to 20;
        rand_num=ranuni(54877);
        output;
end;
run;

proc sort data=randomize;
by rand_num;
run;

data randomize;
set randomize;
temp = 0;
if _n_ >= 6 then temp=1;
if _n_ >= 11 then temp =2;
if _n_ >=16 then temp = 3;
run;
```

```
proc sort data=randomize; by bag; run;

PROC FORMAT;
  VALUE  fortemp 0="Low"
                 1="Med"
                 2="High"
                 3="Very High";
run;

proc print data=randomize;
FORMAT  temp fortemp.;
var bag temp;
run;
```

1	1	Med
2	2	High
3	3	Med
4	4	Low
5	5	Very High
6	6	Low
7	7	Low
8	8	High
9	9	High
10	10	Very High
11	11	Very High
12	12	Med
13	13	High
14	14	Very High
15	15	High
16	16	Low
17	17	Med
18	18	Med
19	19	Very High
20	20	Low

Block Randomization

- Block randomization is a tool to eliminate the effect of possible confounders through design.
- Vitamin C example: we are interested in the reduction in the number of colds when children take vitamin C. We randomly assign treatment (vitamin C or not) to children, and record the number of colds. Independent variable? Dependent variable? Experimental units?
    - Experimental unit: child.
    - Treatments: taking vitamin C or not. X_{i1}
    - Outcome is the number of colds.
- Fit the model $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$.
- For equal group sizes, $var(b_1) = \frac{\sigma^2}{\sum(X_{i1} - \overline{X})^2} = \frac{4\sigma^2}{n}$
- For a fixed n, how can we reduce the variance?
    - Think of where the variance comes from?
    - Maybe males and females differ in the amount of colds they get in general $\Rightarrow$ accounting for sex would reduce σ^2.

Block Randomization

- Block randomization is a tool to eliminate the effect of possible confounders through design.
- Vitamin C example: we are interested in the reduction in the number of colds when children take vitamin C. We randomly assign treatment (vitamin C or not) to children, and record the number of colds. Independent variable? Dependent variable? Experimental units?
  - Experimental unit: child.
  - Treatments: taking vitamin C or not. X_{i1}
  - Outcome is the number of colds.
- Fit the model $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$.
- For equal group sizes, $var(b_1) = \frac{\sigma^2}{\sum(X_{i1} - \overline{X})^2} = \frac{4\sigma^2}{n}$
- For a fixed n, how can we reduce the variance?
  - Think of where the variance comes from?
  - Maybe males and females differ in the amount of colds they get in general $\Rightarrow$ accounting for sex would reduce σ^2.

Block Randomization (cont.)

- Separate the data in different blocks.
  - These are the homogenous groups.
  - Males and females in our example.
- Randomize inside each block.
- Why do this? Wouldn't complete randomization work? Consider extreme case when $n = 4$.
  - With large samples, about 50% of males and females should receive vitamin C.
  - In practice, even if you work for Google, n is ALWAYS limited.
  - Are not guaranteed to have these groups balanced unless you do block randomization.

Six Popular Designs

1. Completely Randomized Design
2. Factorial Experiments
3. Randomized Complete Block Designs
4. Nested Designs
5. Repeated Measures Designs
6. Incomplete Bock Designs

Completely Randomized Design

- The simplest of all designs.
- You have c treatment groups.
- Each experimental subject is assigned to a treatment with an equal probability.
- Y_{ij} is the j^{th} replicate for the i^{th} treatment group.
- Analysis of variance model:
  - $Y_{ij} = \beta_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma^2)$
  - Test hypotheses regarding β_i's.
- Ch 16-18 in KNNL.

Factorial Experiments

- Consider completely randomized designs for multi-factor studies.
- These are called factorial designs.
  - If we have 3 factors with levels f_1, f_2 and f_3
  - have $f_1 \times f_2 \times f_3$ treatment groups.
  - Called an $f_1 \times f_2 \times f_3$ design.
- Analysis of variance model can still hold.
  - Usually write it as main and interaction effects.
  - Testing of interactions is usually of interest.
  - Also have "hidden replication". Useful when all combinations are only collected once.
- Ch 19 and 24

Randomized Complete Block Designs

- You are interested in one factor.
- There is imbalance across another factor.
- Randomize within this other factor (Blocking).
- Consider the vitamin C example (block was sex, which had 2 levels). Fit:
  - $Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \epsilon_{ij}$
  - Y_{ij} is the number of colds for i^{th} child in block j
  - X_{ij1} is an indicator for treatment.
  - X_{ij2} is an indicator for block.
- Would want to test $H_0 : \beta_1 = 0$.
- Ch 21

Nested Designs

- Nested designs differ from cross designs (or factorial designs) in that:
  - certain levels of one factor can only occur for levels of another factor.
- Recall the operator example.
  - Operators 1,2,3 only work at plant 1.
  - Operators 4,5,6 only work at plant 2, etc.
- We typically analyze these using a mixed effects model. Will talk about this later...
- Ch 26

Incomplete Block Designs

- Block sizes are smaller than the number of treatments.
- Used when you physically cannot do all of the experiments on each block.
    - Often when the block is a subject.
    - Cannot have a person try all 36 types of ice cream a store makes.
    - Have all subjects taste a subset of possibilities.
- Ch 28

Mixed-Effects Models:repeated measures design

- Conditional on factors, your data are correlated.
- Ignoring correlation is common, and leads to spurious results.
- Example:
  - We use HRV to measure the balance of the sympathetic to parasympathetic nervous system that is associated with acute stress.
  - We enroll 100 Ph.D. students in the study.
  - The students are from different schools.
  - We tell 50 students that they have funding next year and 50 that they do not.
  - We measure this score every night for 30 days after finding out the news.
  - How does stress change over time?
  - Observations from the same subject are correlated.
  - Observations from the same school are correlated.

- $Y_{ijkt} = (\beta_0 + \beta_1 t) + (\beta_2 + \beta_3 t)\, I(j = 0) + \alpha_i + \gamma_{ijk} + \epsilon_{ijkt}$
- kth student in the ith school that has $j = 1$ if they have money and $j = 0$ otherwise.
- α_i are random effects for school, and γ_{ijk} is the random effect for subject.
- ϵ_{ijkt} is correlated with $\epsilon_{ijkt'}$.
- Chapter 25. We will look at this in more depth than the book.

Observational Studies

- There are three main classes of observation studies.
- Cross-Sectional Studies look at a single time interval.
  - Look at public records for the number of homicides in Philadelphia in 2006 and if the victim lived near an alcohol seller.
- Prospective studies have the treatment precede the response.
  - Students select if they want to take vitamin C throughout the year or not.
  - You record how many colds they receive in the year.
  - Could have confounders, but temporal relationship helps in terms of possible causality.

- Retrospective studies select experimental units based on outcome.
    - Look at Pittsburghers who were diagnosed with lung cancer in 1975.
    - Look at Pittsburghers who were not diagnosed with lung cancer in 1975.
    - Compare if they worked in a Steel Mill or not.
    - Often used if an event is rare.