Topics in Research Design and Quantitative Analysis

Alex Sox-Harris, Ph.D., MS
Associate Professor, Department of Surgery, Stanford School of Medicine
Stanford – Surgery Policy Improvement Research and Education (S-SPIRE)
Research Career Scientist, Center for Innovation to Implementation (Ci2i), VA Palo Alto Healthcare System
Goals

- Encourage humility
- Encourage skepticism
- Discuss issues that come up repeatedly in design and statistical consultations.
- Time for general or specific questions
HA! NAME ONE THING THAT YOU DO BETTER THAN ME.

HUMILITY.
Skepticism
Common Consultation Topics

- Equivalence vs Different Hypotheses
- Dependent data
- Sample size determinations
  - Power and precision analyses
- Multiple comparisons or tests and alpha adjustments
- Quasi-experimental methods
- When to use non-parametric methods
- Machine learning methods
Equivalence Studies

Researchers often want to evaluate if a new intervention is equivalent to an existing intervention in terms of complications or outcomes.

The equivalence of two interventions cannot be established by failing to find a statistical difference between them!

To assess the difference between interventions. You are interested in finding a difference.

- Null Hypothesis: \( \text{Mean 1 - Mean 2} = 0 \)
- Alternative Hypothesis: \( \text{Mean 1 - Mean 2} \neq 0 \)

Power Analysis: Need to specify the smallest difference that would be clinically meaningful (Effect Size).

Analysis: Independent sample t-test

\( p \)-value is the probability of the results given the null hypothesis is true.

Does the 95% CI for Mean 1 - Mean 2 include zero?
Difference Trial Example

- Procedure 1 is the standard of care. You think Procedure 2 can improve outcomes as measured by the Surgical Outcome Measure (SOM).

- Power Analysis:
  - Historically, Procedure 1 has resulted in scores with a mean = 50 and an SD = 10.
  - You think that an improvement of 5 points is clinically meaningful and you are willing to assume the SD will also be 10 with Procedure 2.
  - This translates into a standardized effect size of $\frac{5}{10} = 0.5$. Stipulating an alpha = .05, and power = .80.
Difference Trial Example

- Running the power analysis gives you this:
  - Group sample sizes of 64 and 64 achieve 80% power to detect an SMD of .50
  - Significance level (alpha) of 0.05 using a two-sided two-sample t-test.
Results

Mean (SD) of Procedure 1 50.4 (10.2)
Mean (SD) of Procedure 2 53.3 (14.2)

t = 1.31, p-value = 0.19

95 percent confidence interval of M2-M1: [-1.5 to  7.2]
Why can’t you say the procedures are the same if you fail to reject the null with a difference test (e.g., t-test)?

- Because “same” has not been defined or included in the analysis
- Because values in the CI that would fail to reject the null may contain values that are clinically meaningful!!
95% CI for Mean Difference Proc1 vs.
Equivalence or Non-superiority Trial

- Goal: To assess if interventions are clinically equivalent.
- Evidence that they are equivalent would be meaningful because the new procedure has other benefits such as cost or safety.
- Often, the one-sided version of this design is used (the non-superiority trial) to assess if the new procedure is “at least as good as” the old procedure.
Equivalence Trial

- **Null Hypothesis:**
  \[ |\text{Mean 1} - \text{Mean 2}| \geq \Gamma \text{ (gamma), a pre-specified threshold below which is “clinically meaningless”} \]

- **Alternative Hypothesis:**
  \[ |\text{Mean 1} - \text{Mean 2}| < \Gamma \]
Equivalence or Non-superiority Trial

- Power Analysis: Need to specify the biggest difference that would be clinically meaningless (Effect Size).
- Using the example from above, if 5 SOM points is clinically meaningful, then presumably the threshold for clinically meaningless is less than 5.
- Let's say that we decide that a difference of 2 SOM points is basically meaningless. So the null hypothesis is the $|\text{Mean } 1 - \text{Mean } 2| \geq 2$. 
Power Analysis

- sample sizes of 226 in the first group and 226 in the second group achieve 80% power at a 0.10 one-sided significance level. (overall alpha = 0.05)
Results

Mean (SD) of Procedure 1  55.2 (9.7)
Mean (SD) of Procedure 2  55.6 (9.2)

90 percent confidence interval for the mean difference: [-1.9 to 1.0]

TOST procedure (two one sided tests): p = 0.04
Figure 1: Interpretation of 4 Confidence Intervals for a Difference Trial (H0: M1-M2=0; HA: M1-M2≠0) and an Equivalence Trial (H0: |M1-M2|>10; HA: |M1-M2|≤10)

**Difference Trial (α=.05):**
1. No difference
2. Difference
3. No difference
4. Difference

**Equivalence Trial (d=10; α=.025):**
1. Not equivalent
2. Not equivalent
3. Equivalent
4. Equivalent

Mean Difference of Treatment 1-Treatment 2
“Not Statistically Different” Does Not Necessarily Mean “the Same”: The Important but Underappreciated Distinction Between Difference and Equivalence Studies

Alex H.S. Harris, MS, PhD, Sara Fernandes-Taylor, PhD, and Nicholas Giori, MD, PhD

Researchers often want to evaluate whether a new medical or surgical treatment is equivalent to an existing treatment. The new treatment may be preferred if its results are equivalent to those of the existing approach in terms of complications or outcomes but it is superior in terms of ease of use, safety, or cost. However, many researchers are unaware that the equivalence specifically type-I and type-II errors. A type-I error, or false positive, is an error in which the null hypothesis is rejected when, in fact, the null hypothesis is true. For example, a type-I error has occurred if researchers declare that two treatments produce different outcomes when, in reality, no difference exists. The probability of a type-I error is denoted by $\alpha$. A type-II
Statistics 101 only covers methods that have a strong assumption of independent errors (e.g., ANOVA, independent sample t-tests, OLS regression).

Many of our data and questions have dependencies that require other less familiar methods.

Dependent, non-independent, correlated, nested, clustered errors….All the same thing.
Goals

• Be able to recognize dependences in data.
  • Patients within clinics
  • Repeated measures on units
  • Longitudinal data
• Understand dangers of ignoring this issue
• Highlight common bad methods
• Provide a basic orientation to one statistical framework for handling dependencies: Mixed-effects regression
Common Data Structures

- Multi-Level Organizational Data
  - Patients within providers within facilities
Common Data Structures

- Repeated measures per unit
  - Several BPs per person at each assessment and/or over time
  - Several assays per culture
Common Data Structures

- Repeated-measures on individuals over time
  - Monthly measurement of disease status

![Graph of repeated measures over time]
Common Data Structures

- Both within person and within organization clustering
  - RCT where providers are the unit of randomization
  - Outcomes are patient-level trajectories
  - Site\provider\patient\
The Problem

- Common statistical tools have no good way of dealing with multi-level details (correlated errors, sample size, variances)
  - OLS Regression
  - ANOVA
  - t-tests
- It matters – failing to attend to these details can give very wrong results.
Old (and usually bad) Solutions

- To aggregate or disaggregate data to one level and apply familiar statistical models.
Example

- **Study:** What are the clinic characteristics (e.g., co-located social work service) that influence patient outcomes?
  - Sample is 700 patients in 20 clinics

- **Bad Solutions:**
  - Analyze all information at the patient-level
  - Force all information to the clinic-level
Usual Methods Get This Wrong

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Clinic ID</th>
<th>Patient Outcome</th>
<th>Clinic Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>etc</td>
<td>etc</td>
<td>etc</td>
<td>etc</td>
</tr>
</tbody>
</table>
Forcing Information to the Patient-level

- Confounds patient and clinic sample sizes
- Radically reduces the SE of parameter estimates
- Leads to more null-hypothesis rejection and inappropriately narrow CIs
Force all information to the Clinic-level

<table>
<thead>
<tr>
<th>Site ID</th>
<th>Patient Outcome</th>
<th>Site Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.5</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5.6</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>8.2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>9.7</td>
<td>1</td>
</tr>
</tbody>
</table>
Force all information to the Clinic-level

- Lose power
- Lose information about within clinic variability and sample size
Compare Methods

- OLS regression on 700 observations
- OLS regression on 20 observations
- Mixed-Effects Regression
- Test t statistic
  - 10.0, 3.2, 3.5
Mixed Effects Regression (HLM, mixed models, random effects models, etc.)

- Keeps track of multi-level details and allows for dependencies.
- Generalized versions (logistic, Cox, count)
- Handles unbalanced data and variable assessment schedules, all cases can be included
- Implemented in most major packages
- Other strategies/models are available that handle some of these details (robust/shrunken SE; fixed effects models; GEE).
Mixed Effects Regression

- Address single-level questions while accounting for dependencies at other levels.
  - Do patients who have a particular procedure have better outcomes?
- Test interesting and important multi-level (cross-level) hypotheses.
  - Does surgical setting (e.g., academic, private group practice) impact procedural effectiveness on patient outcomes?
Real Life Example

- What are the factors that are associated with surgeons requesting general anesthesia for carpal tunnel release?
- 16,053 procedures
- 785 surgeons
- 111 VA hospitals
glm(formula = ga ~ c.age + gender.factor + race + marital.rec + serviceConnect.factor + asaClass.factor, family = binomial, data = x)

Deviance Residuals:
  Min       1Q   Median       3Q      Max
-0.8106  -0.5965  -0.5631  -0.5228   2.1832

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.945842   0.196584  -9.898  < 2e-16 ***
c.age       -0.008454   0.002006  -4.215 2.50e-05 ***
gender.factorM -0.152158   0.070464  -2.159   0.0308 *
raceNon-Hispanic Black 0.297721   0.061571   4.835 1.33e-06 ***
raceOther minority   0.061751   0.081951   0.754   0.4511
marital.recDivorced/separated 0.058132   0.050464   1.152   0.2493
marital.recNEVER MARRIED   0.067604   0.082320   0.821   0.4115
marital.recSingle or widow/widower -0.089112  0.109629  -0.813   0.4163
serviceConnect.factorNSC or <50%  -0.071935  0.045883  -1.568   0.1169
serviceConnect.factorOther  -0.189297  0.200324  -0.945   0.3447
asaClass.factor2          0.260787   0.185701   1.404   0.1602
asaClass.factor3          0.417518   0.186095   2.244   0.0249 *
asaClass.factor4          0.578912   0.230124   2.516   0.0119 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
Family: binomial  ( logit )
Formula: ga ~ c.age + gender.factor + race + marital.rec + serviceConnect.factor + asaClass.factor + (1 | sta3n.factor/surgeon.factor)

Random effects:
Groups                      Name        Variance Std.Dev.
surgeon.factor:sta3n.factor (Intercept) 1.904    1.380
sta3n.factor                (Intercept) 3.121    1.767

Number of obs: 16053, groups:  surgeon.factor:sta3n.factor, 785; sta3n.factor, 111

Fixed effects:

  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -2.658075   0.308538  -8.615  < 2e-16 ***
c.age                  -0.013861   0.002591  -5.349 8.86e-08 ***
gender.factorM        -0.038550   0.089761  -0.429    0.668
raceNon-Hispanic Black  0.078207   0.083720   0.934    0.350
raceOther minority    -0.157742   0.108698  -1.451    0.147
marital.recDivorced/separated  0.100191   0.063846   1.569    0.117
marital.recNEVER MARRIED  0.026475   0.104431   0.254    0.800
marital.recSingle or widow/widower -0.028766   0.136205  -0.211    0.833
serviceConnect.factorNSC or <50%   0.070086   0.058559   1.197    0.231
serviceConnect.factorOther -0.055370   0.254383  -0.218    0.828
asaClass.factor2      -0.019916   0.230841  -0.086    0.931
asaClass.factor3     -0.149605   0.232928   0.642    0.521
asaClass.factor4     -0.150451   0.291095   0.517    0.605
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Suggested Resources for Learning About Mixed Effects Models

- Hox, J. J. *Applied Multilevel Analysis*. 
  [http://www.fss.uu.nl/ms/jh/publist/amaboek.pdf](http://www.fss.uu.nl/ms/jh/publist/amaboek.pdf)
Multi-level thinking is powerful and important conceptually and statistically.

Need to use models that keep track of multi-level details:
- Sample size
- Variance partition
- Correlated errors

Unless you want to commit a lot of time to this, ask for help.
Consultation Mistakes to Avoid

- Last minute requests for meetings/analyses
- Relying on too much on email, especially in lieu of an initial meetings
- Unclear expectations regarding effort, authorship, credit
- Things that make statisticians heads explode:
  - Power analyses after a study is done
  - Messy datasets
  - Requests for “quick” analyses
Time Allowing…

- Power Analysis
- Alpha adjustments for multiple tests
- Parametric methods vs. alternatives
- Anything else?
Thank You

alexsox@stanford.edu