Simulation Methods in Epidemiologic Research and Learning

Matthew Fox
Department of Epidemiology
Center for Global Health and Development
Boston University, USA
Random Error and 95% CIs

- If you ask most people, a 95% confidence interval from 1.1 to 2.3 means:
  - There is a 95% chance that the true value is between 1.1 and 2.3
  - This is not correct

- If statistical model is correct and no bias, a confidence interval derived from a valid test statistic will, over unlimited repetitions of the study, contain the true parameter with a frequency no less than its confidence level (e.g. 95%)
  - Simple simulation helps make the distinction
Simulate the height of 1000 people with a mean of 65 and std of 5

From the initial 1000, simulate 1000 datasets each drawn from the original of size 20 and for each calculate a mean and 95% CI
How Often Did CI Contain the Truth?

Full sample

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>65.3225048</td>
<td>4.9252091</td>
<td>50.7579163</td>
<td>86.5469094</td>
</tr>
</tbody>
</table>

Did the 95% CI include the true value?

<table>
<thead>
<tr>
<th>included</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>53</td>
<td>5.30</td>
<td>53</td>
<td>5.30</td>
</tr>
<tr>
<td>Yes</td>
<td>947</td>
<td>94.70</td>
<td>1000</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Outline

- How SimPLE started
- What we’ve done
- How you can do it
- Some examples
- Why it is important
DISCLAIMER:

I am not an expert in data simulations ... and this is the point!
A Useful SAS Book

Simulating Data with SAS

Rick Wicklin
Motivation

- In my doctoral program I was always wanting a “confounded” dataset when TAing or getting ready for exams, yet at first I didn’t know how to create one
  - Found out that in order to simulate it, you have to understand it well enough
  - Started to realize what I didn’t know
  - Started to realize I could figure out things myself
- I had a colleague who said that he took a class in which for every concept they learned, they had to simulate a dataset that illustrated that problem
Epi Doctoral Qualifier Question

Below is a shell table for a dataset on the relationship between an exposure E and an outcome D stratified by a covariate C. Assume that we could know each person in the study’s counterfactual susceptibility type (Type 1-4)*. Create a dataset with the following properties and fill in the table below:

1. The crude E-D relationship is confounded by C (by statistical criteria)
2. The C stratum-specific estimates of the E-D relationship are unconfounded (by statistical criteria)
3. \( P1 \) is not equal to \( Q1 \)*
4. There is no effect measure modification by C of the ED relationship on the difference scale but there is effect measure modification on the relative scale

*Greenland S, Robins J Identifiability, Exchangeability, and Epidemiological Confounding *IJE 1986; 15: 413-419
So Was the Birth of SimPLE

- **SIMulating Problems for Learning Epidemiology**
- **Goals:**
  - Bring together doctoral students from epidemiology and environmental health to learn
  - Everyone contributes
  - We are all beginners
  - We all choose a topic to try to understand better

- **Took us a few sessions to cover some very simple concepts and everyone was off and running**
  - Message: basic simulation for learning is not hard to do!
What Have We Covered

- Simulating datasets
- Simulating datasets with particular structures
  - Confounding, collider bias, effect measure modification
- Simulating dataset from the main dataset with bias
  - Selection bias, measurement error
- Understanding M bias
- Quantitative bias analysis
- Dependent error
- Bootstrapping
What Do I Consider a Simulation?

- Often we think of big scary, hairy simulations with lots of parameters to vary, complex error structures, lots of complex formulas and always done by a biostatistician
- I consider everything from
  - Demonstration of a concept
  - Creation of a static toy dataset with no randomness
  - Creation of a dataset based on probabilities
  - Varying parameters
  - Simulating error, and error structures
  - Big hairy simulations with lots of variation
Simple Simulations
Simulate an Exact Dataset

- data summary;
  - input exp out count;
  - cards;
  - 1 1 25
  - 1 0 75
  - 0 1 50
  - 0 0 50
  - ;
- run;
- proc freq data=summary;
  - tables exp*dis/nocol nopercent;
  - weight count;
- run;
Simulate an Exact Individual Level Dataset

- Create the 2x2 table
- data individual;
  - do j = 1 to 25;
    - exp = 1; dis = 1; output;
  - end;
  - do j = 1 to 75;
    - exp = 1; dis = 0; output;
  - end;
  - do j = 1 to 50;
    - exp = 0; dis = 1; output;
  - end;
  - do j = 1 to 50;
    - exp = 0; dis = 0; output;
  - end;
- run;

<table>
<thead>
<tr>
<th></th>
<th>E+</th>
<th>E-</th>
</tr>
</thead>
<tbody>
<tr>
<td>D+</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>D-</td>
<td>75</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Random Number Generators

- Often want to draw randomly from a distribution rather than create exact outputs
- SAS has lots of random number generators
  - `RAND('BERNOULLI', probability);`
  - `RANBIN(seed, # trials, probability);`
  - `RANUNI(seed);`
  - `RANTRI(seed, mode)`
  - `RANNOR(seed, x);`
  - and more… see SAS documentation
Simulate a Simple Dataset Probabilistically

- Pr(E+) is 50%
- Pr(D+) is 25% if E-
- Pr(D+) is 50% if E+

data prob;
  - do j = 1 to 10000;
    - exp = rand('bernoulli',0.5);
    - if exp = 0 then dis = rand('bernoulli',0.25);
    - else if exp = 1 then dis = rand('bernoulli',0.5);
  - output;
  - end;
- run;
DAGs to Simulate Data

- There are other ways, for me this is the simplest
- Can simulate from a regression model
- (See book for details)
- Can build complex error structures
Confounding
### N=1000 per stratum
C should be associated with E and D

<table>
<thead>
<tr>
<th></th>
<th>Crude</th>
<th>C-</th>
<th>C+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E+</td>
<td>E-</td>
<td>E+</td>
</tr>
<tr>
<td><strong>D+</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>170</td>
<td>80</td>
</tr>
<tr>
<td><strong>D-</strong></td>
<td>840</td>
<td>830</td>
<td>120</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1000</td>
<td>1000</td>
<td>200</td>
</tr>
<tr>
<td><strong>Risk</strong></td>
<td>0.16</td>
<td>0.17</td>
<td>0.4</td>
</tr>
<tr>
<td><strong>RR</strong></td>
<td>0.94</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

\[
\text{RR}_{CD|E-} = 4 = \frac{0.2}{0.05}
\]

\[
\text{RR}_{CE} = 4 = \frac{800/1000}{200/1000}
\]
Simulating DAGs: Confounding

- Define the baseline risks
  - What % of people have C+?
  - What % of people C- are E+
  - What % of people C- and E- are D+

- Define effects (relative vs absolute)
  - What is the RR/RD for C on E?
  - What is the RR/RD for C on D?
  - What is the RR/RD for E on D?

- Define interactions
  - Do E and C interact to cause D?
  - If so, on what scale?

---

**Simulation Studies for Epidemiology**

- \( \text{Pr}(C+ = 0.5) \)
- \( \text{RR}_{CE} = 2.5 \)
- \( \text{RR}_{CD} = 2 \)
- \( \text{RR}_{ED} = 5 \)
- \( \text{Pr}(E+|C- = 0.15) \)
- \( \text{Pr}(D+|C-,E- = 0.05) \)
Simulate Confounding Probabilistically

- data = conf;
- do j = 1 to 10000; * sample size;
  - conf = rand('bernoulli', 0.5); * sim confounder;
  - if conf = 1 then exp = rand('bernoulli', 0.15*2.5); * E|C+;
  - else if conf = 0 then exp = rand('bernoulli', 0.15); * E|C-
  - if exp = 1 and conf = 1 then dis = rand('bernoulli', 0.05*2*5);
  - else if exp = 1 and conf = 0 then dis = rand('bernoulli', 0.05*5);
  - else if exp = 0 and conf = 1 then dis = rand('bernoulli', 0.05*2);
  - else if exp = 0 and conf = 0 then dis = rand('bernoulli', 0.05);
- output;
- run;

Table of exp by dis

<table>
<thead>
<tr>
<th>exp</th>
<th>dis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>1068</td>
<td>1504</td>
</tr>
<tr>
<td></td>
<td>41.52</td>
<td>58.48</td>
</tr>
<tr>
<td>-</td>
<td>511</td>
<td>6917</td>
</tr>
<tr>
<td></td>
<td>6.88</td>
<td>93.12</td>
</tr>
<tr>
<td>Total</td>
<td>1579</td>
<td>8421</td>
</tr>
</tbody>
</table>

Statistics for Table of exp by dis

Estimates of the Relative Risk (Row1/Row2)

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Value</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control (Odds Ratio)</td>
<td>9.6121</td>
<td>8.5315 10.8297</td>
</tr>
<tr>
<td>Cohort (Coll Risk)</td>
<td>6.0360</td>
<td>5.4867 6.6403</td>
</tr>
</tbody>
</table>

Estimates of the Common Relative Risk (Row1/Row2)

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Method</th>
<th>Value</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control (Odds Ratio)</td>
<td>Mantel-Haenszel</td>
<td>8.2128</td>
<td>7.2595 9.2913</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>7.9946</td>
<td>7.0651 9.0463</td>
</tr>
<tr>
<td>Cohort (Coll Risk)</td>
<td>Mantel-Haenszel</td>
<td>5.0930</td>
<td>4.6082 5.6289</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>5.0643</td>
<td>4.5876 5.5906</td>
</tr>
</tbody>
</table>
Simulating DAGs

- Find the independent nodes and simulate
  - Specify probability
- Simulate nodes dependent on one arrow
  - Specify probability in all levels of the arrows the leads into the node
- Simulate nodes dependent on only two arrows, etc.
  - Specify probability in all levels of arrows that lead into the node
- Pay attention to scale, additive or multiplicative
- Pay attention to interaction (additive or multiplicative)
Unmeasured Confounders

- Suppose I have data on E and D and want to simulate U?
- Now the E and D variables exist, can’t simulate E and D dependent on U and C
- Instead I need to simulate U based on the probability of being in any of the 8 missing cells in the table
  - $\text{RR}_{UD} = 2.5$, $\Pr(U+|E+) = 10\%$ $\Pr(U+|E-) = 20\%$

<table>
<thead>
<tr>
<th></th>
<th>Crude</th>
<th>U+</th>
<th>U-</th>
</tr>
</thead>
<tbody>
<tr>
<td>E+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D+</td>
<td>a 45</td>
<td>E+</td>
<td>E-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1</td>
<td>B1</td>
</tr>
<tr>
<td>D-</td>
<td>c 255</td>
<td>d 630</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C1</td>
<td>D1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Crude</th>
<th>U+</th>
<th>U-</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D+</td>
<td>a 45</td>
<td>E+</td>
<td>E-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A0</td>
<td>B0</td>
</tr>
<tr>
<td>D-</td>
<td>c 255</td>
<td>d 630</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>C0</td>
<td>D0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th></th>
<th>Total</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>m 300</td>
<td>n 700</td>
<td>M1 30</td>
<td>N1 140</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M0 270</td>
<td>N0 560</td>
</tr>
</tbody>
</table>

Diagram:
- U
- E
- D
## Unmeasured Confounders

- \( \text{RR}_{CD} = 2.5 \) and 
  
  \[
  A_1 = \frac{\text{RR}_{CD} M_1 a}{\text{RR}_{CD} M_1 + m - M_1} \quad A_1 = \frac{2.5 \cdot 30 \cdot 45}{2.5 \cdot 30 + 300 - 30}
  \]
  
  \[
  B_1 = \frac{\text{RR}_{CD} N_1 b}{\text{RR}_{CD} N_1 + n - N_1} \quad B_1 = \frac{2.5 \cdot 140 \cdot 70}{2.5 \cdot 140 + 700 - 140}
  \]

- So \( A_1 = 9.8 \) and \( B_1 = 26.9 \)
- And we can now fill in the rest of the table

<table>
<thead>
<tr>
<th></th>
<th>Crude</th>
<th>U+</th>
<th>U-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E+</td>
<td>E-</td>
<td>E+</td>
</tr>
<tr>
<td>D+</td>
<td>a 45</td>
<td>b 70</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B1</td>
</tr>
<tr>
<td>D-</td>
<td>c 255</td>
<td>d 630</td>
<td>C1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>D1</td>
</tr>
<tr>
<td>Total</td>
<td>m 300</td>
<td>n 700</td>
<td>M1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>N1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>M0</td>
</tr>
</tbody>
</table>
Unmeasured Confounders

- So now for any person, if I know their E and D I can tell you the probability of having U:
  - \( \Pr(U+|E+,D+) = \frac{9.8}{45}, \Pr(U+|E+,D-) = \frac{20.2}{255} \)
  - \( \Pr(U+|E-,D+) = \frac{26.9}{70}, \Pr(U+|E-,D-) = \frac{113.1}{630} \)

- Code:
  - if \( E=1 \) and \( D=1 \) then \( U = \text{rand('bernoulli', 9.8/45)} \);
  - else if \( E=1 \) and \( D=0 \) then \( U = \text{rand('bernoulli', 20.2/255)} \);
  - else if \( E=0 \) and \( D=1 \) then \( U = \text{rand('bernoulli', 26.9/70)} \);
  - else if \( E=0 \) and \( D=0 \) then \( U = \text{rand('bernoulli', 113.1/630)} \);

<table>
<thead>
<tr>
<th>Crude</th>
<th>E+</th>
<th>E-</th>
<th>U+</th>
<th>E+</th>
<th>E-</th>
<th>U-</th>
<th>E+</th>
<th>E-</th>
</tr>
</thead>
<tbody>
<tr>
<td>D+</td>
<td>a</td>
<td>b</td>
<td>A1</td>
<td>B1</td>
<td></td>
<td></td>
<td>A0</td>
<td>B0</td>
</tr>
<tr>
<td>D-</td>
<td>c</td>
<td>d</td>
<td>20.2</td>
<td></td>
<td></td>
<td></td>
<td>35.2</td>
<td>43.1</td>
</tr>
<tr>
<td>Total</td>
<td>m</td>
<td>n</td>
<td>30</td>
<td>140</td>
<td></td>
<td></td>
<td>270</td>
<td>560</td>
</tr>
</tbody>
</table>

Simulation Studies for Epidemiology
Three Posters Here at SER

- **100-S Implications of Nondifferential Dependent Misclassification of Covariate and Exposure**
  - Kelly Getz and Alana Brennan
  - TUESDAY, JUNE 24, 2014 7-8:30 PM

- **112-S Understating the Relationship between Directed Acyclic Graphs (DAGs) and Data through Simulation Studies**
  - Julia Rohr
  - TUESDAY, JUNE 24, 2014

- **412-S When Does Adjustment for Predictors of Exposure Misclassification Increase Bias? A Simulation Study**
  - Samantha Parker and Mahsa Yazdy
  - WEDNESDAY, JUNE 25 5:00 – 6:30 pm
Example: Dependent Error

- I had a student whom I asked to simulate dependent error to see when it mattered most
- A colleague had a student who wrote a paper on the same idea (Kelly Getz)
- We brought them together
- SimPLE was born

FIGURE 1. Rearrangement resulting from error in classification of exposure ($E$, $\overline{E}$) and outcome ($D$, $\overline{D}$).