Simulation Methods in Epidemiologic Research and Learning

Matthew Fox
Department of Epidemiology
Center for Global Health and Development
Boston University, USA

BOSTON UNIVERSITY

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

```
data master;
    do j = 1 to 1000;
        seed1=-1;
        x= rannor(seed1);
        height = 65+5*x;
        output;
    end;
    drop j seed1;
run;
```

Simulate the height of 1000 people with a mean of 65 and std of 5

```
proc means data=master n mean std min max; var height;
    title "True population mean height";
run;
```

From the initial 1000, simulate 1000 datasets each drawn from the original of size 20 and for each calculate a mean and 95% CI

```
***************************;
* REPEATED SAMPLING      **;
***************************;
%macro rep();
    %do j = 1 %to 1000;
        data onerep;
            retain seed1;
        * initialize the seed variable;
            if seed1 eq . then seed1 = -1;
        * loop until designated number of controls are found;
                do j = 1 to 20;
                * choose a random person in the database;
                    lookat = round(1000*ranuni(seed1),1)+1;
                * hold that record for the new dataset;
                    set master point=lookat;
                    * output the record to the new dataset if they are eligible;
                    output;
                end;
            drop j seed1;
            stop;
        run;

        proc means data=onerep mean lclm uclm noprint;
            var height; output out=outset lclm=lclm uclm=uclm;
        run;
        data outset; set outset;
            if lclm le 65 le uclm then included=1; else included=0;
            attrib included label="Did the 95% CI include the true value?" format=yn.;
        run;

        proc append base=newset data=outset force; run;
    %end;
```

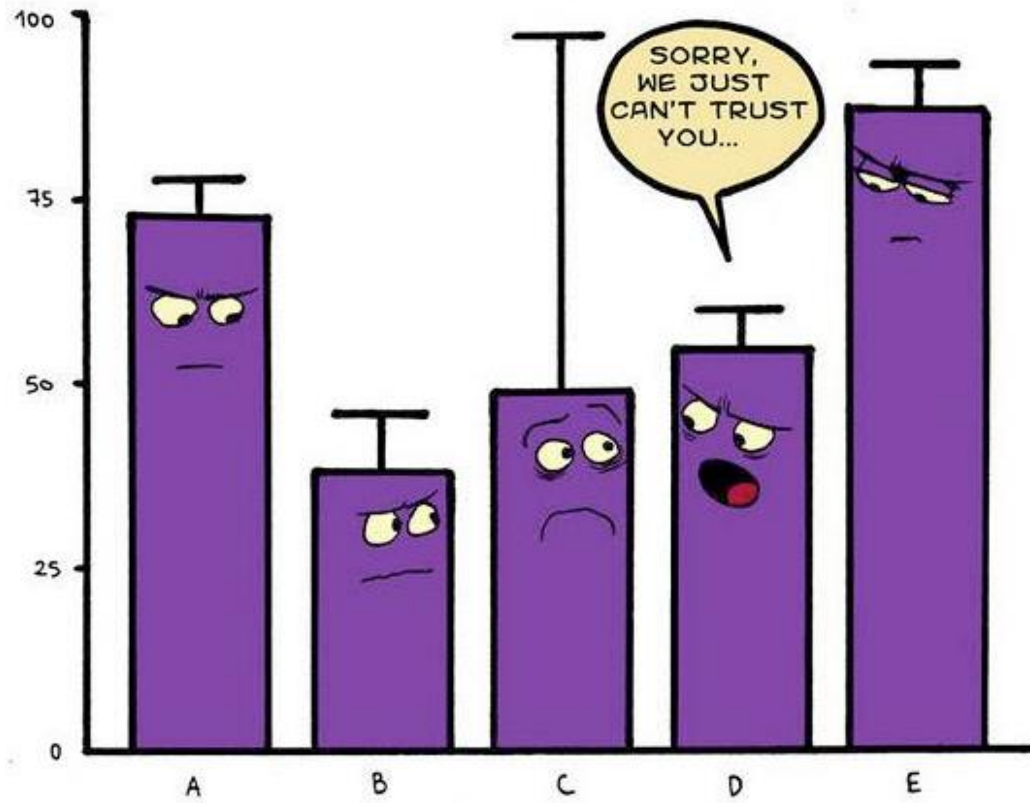
How Often Did CI Contain the Truth?

Full sample

N	Mean	Std Dev	Minimum	Maximum
1000	65.3225048	4.9252091	50.7579163	86.5469094

Did the 95% CI include the true value?

included	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	53	5.30	53	5.30
Yes	947	94.70	1000	100.00



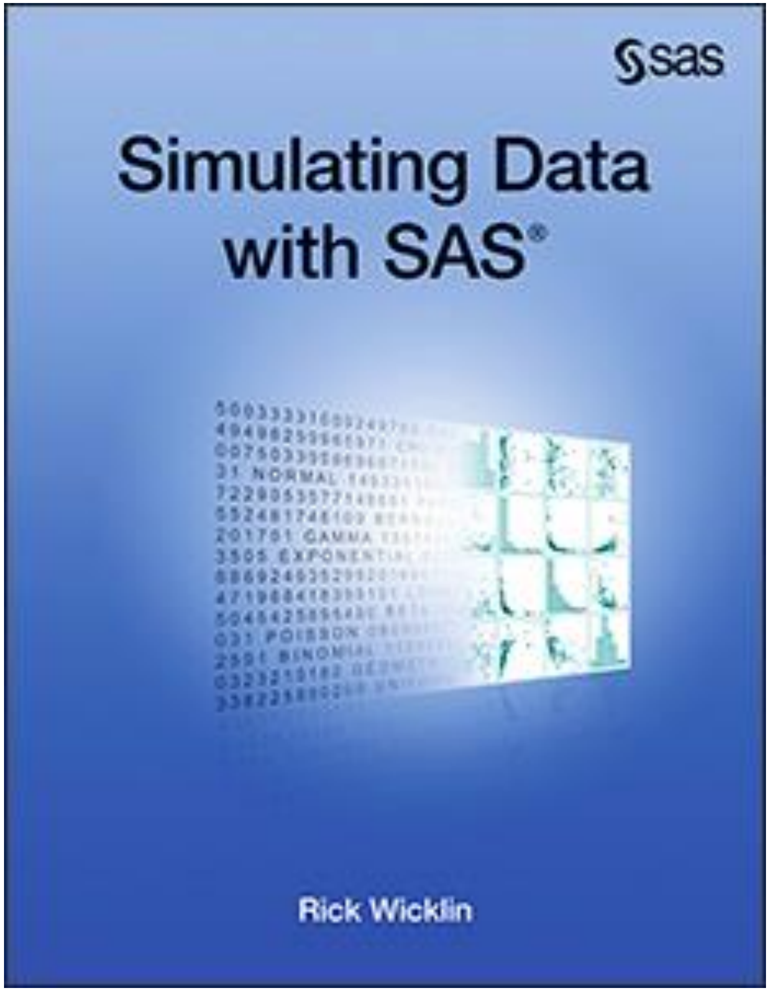
Outline

- 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



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

BU

*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!

BU

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

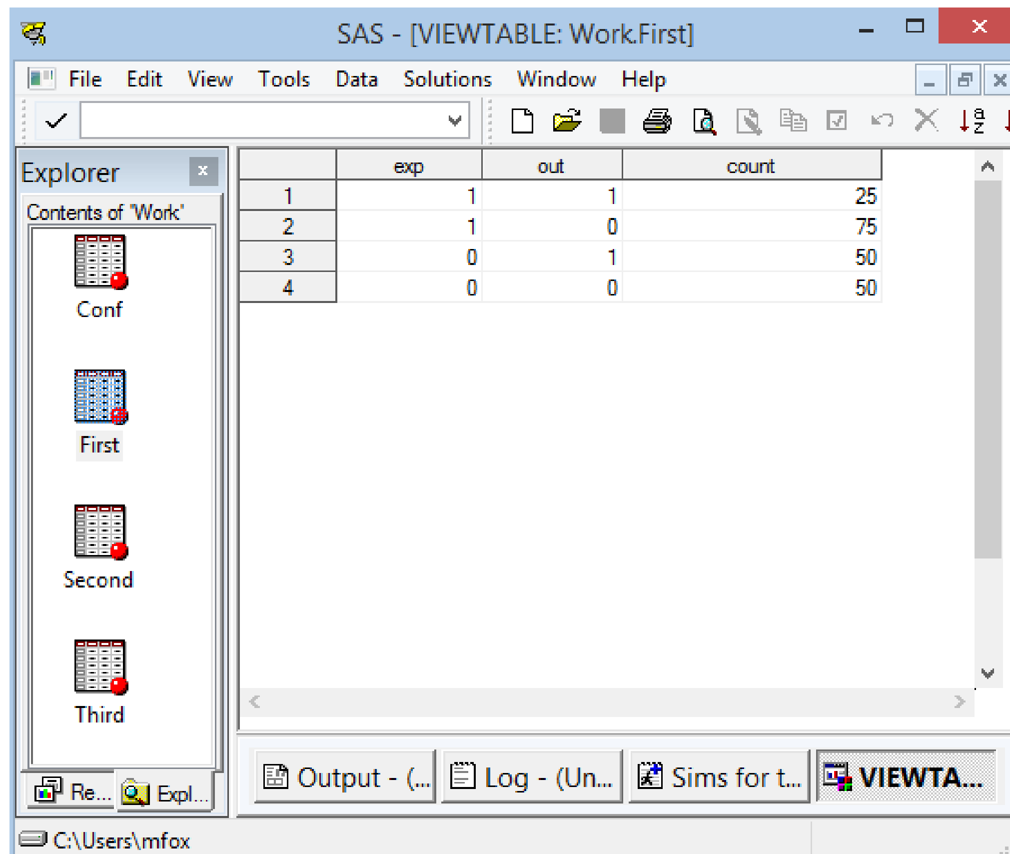
Simple Simulations

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;

Table of exp by dis

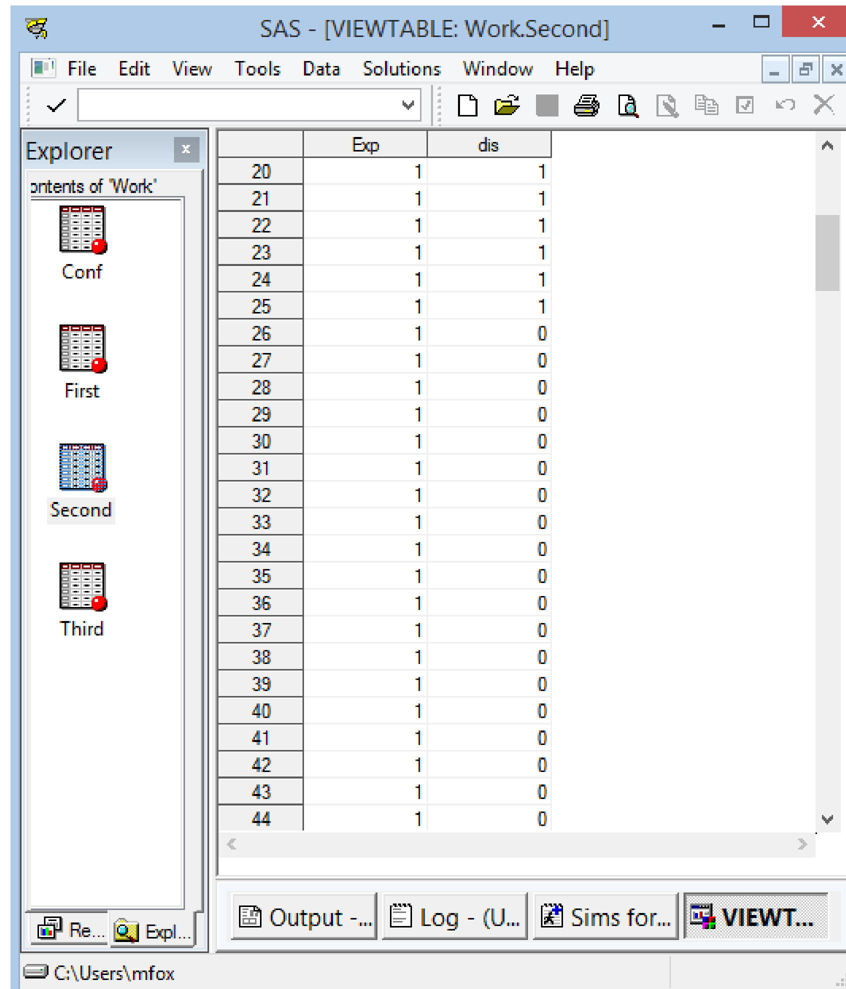
exp / Frequency / Row Pct	dis = 0	dis = 1	Total
0	50 50.00	50 50.00	100
1	75 75.00	25 25.00	100
Total	125	75	200



Simulate an Exact Individual Level Dataset

	E+	E-
D+	25	50
D-	75	50
Total	100	100

- 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;



	Exp	dis
20	1	1
21	1	1
22	1	1
23	1	1
24	1	1
25	1	1
26	1	0
27	1	0
28	1	0
29	1	0
30	1	0
31	1	0
32	1	0
33	1	0
34	1	0
35	1	0
36	1	0
37	1	0
38	1	0
39	1	0
40	1	0
41	1	0
42	1	0
43	1	0
44	1	0

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;

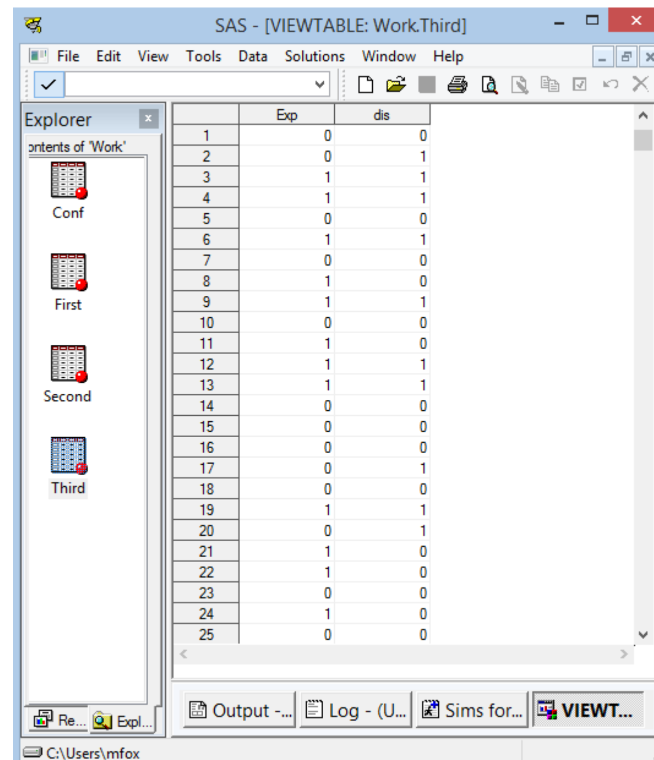
The FREQ Procedure

exp	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	4985	49.85	4985	49.85
1	5015	50.15	10000	100.00

dis	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6328	63.28	6328	63.28
1	3672	36.72	10000	100.00

Table of exp by dis

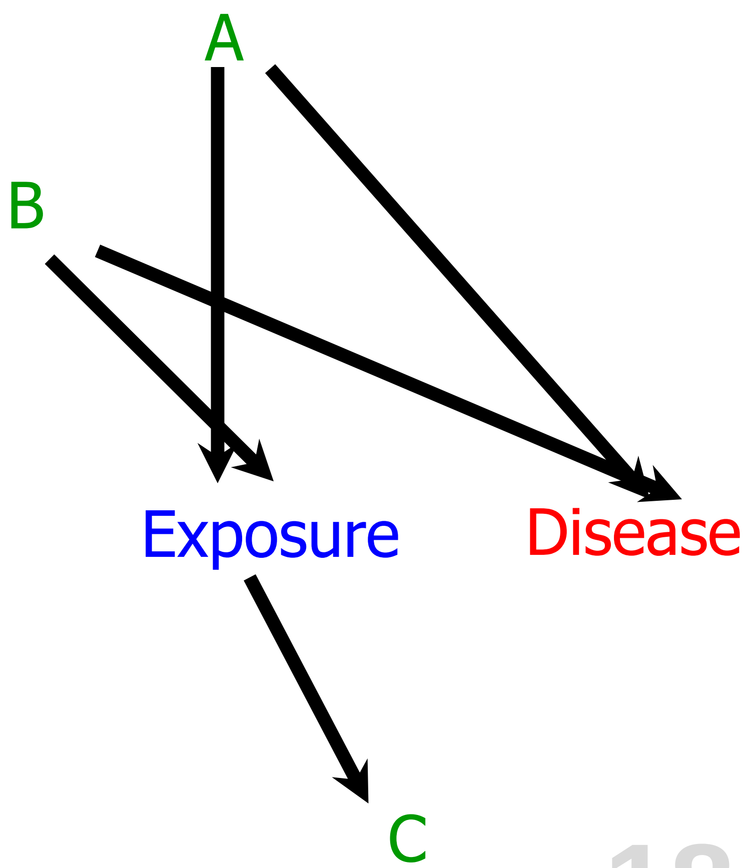
exp / dis (Frequency, Row Pct)	0	1	Total
0	3759 75.41	1226 24.59	4985
1	2569 51.23	2446 48.77	5015
Total	6328	3672	10000



BU

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

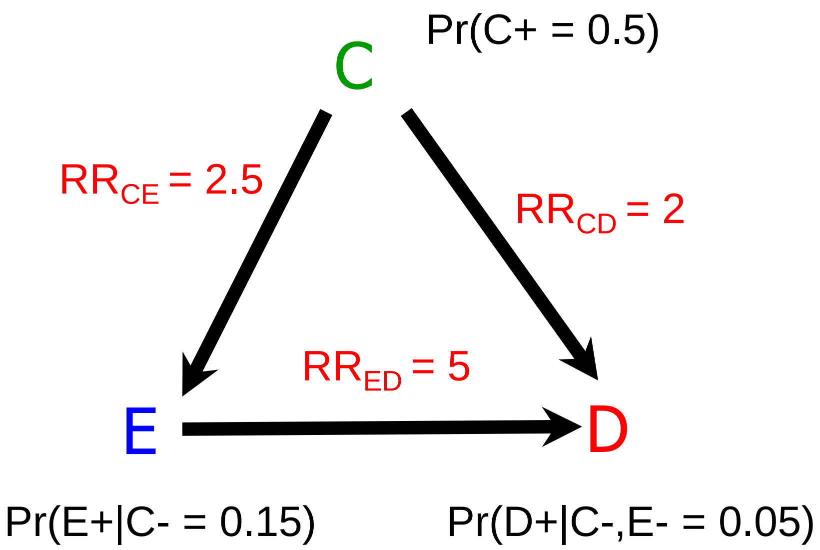
Crude			C-			C+		
	E+	E-		E+	E-		E+	E-
D+	160	170	D+	80	160	D+	80	10
D-	840	830	D-	120	640	D-	720	190
Total	1000	1000	Total	200	800	Total	800	200
Risk	0.16	0.17	Risk	0.4	0.2	Risk	0.1	0.05
RR	0.94		RR	2		RR	2	

$$RR_{CD|E-} = 4 = (0.2/0.05)$$

$$RR_{CE} = 4 = [(800/1000)/(200/1000)]$$

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?



Simulate Confounding Probabilistically

- data conf;
  - d
  -
- run;

SAS - [VIEWTABLE: Work.Conf]

Help

	dis
1	0
0	1
0	0
1	1
0	0
1	0
0	0
0	1
0	0
0	0
0	0
0	0
0	0
1	1
0	0
1	0
1	0
0	0
0	0
1	0
1	0
1	1
0	0
	0
	0

ns for... VIEWT...

Table of exp by dis

exp / dis Frequency Row Pct	+	-	Total
+	1068 41.52	1504 58.48	2572
-	511 6.88	6917 93.12	7428
Total	1579	8421	10000

Statistics for Table of exp by dis

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	9.6121	8.5315	10.8297
Cohort (Col1 Risk)	6.0360	5.4867	6.6403

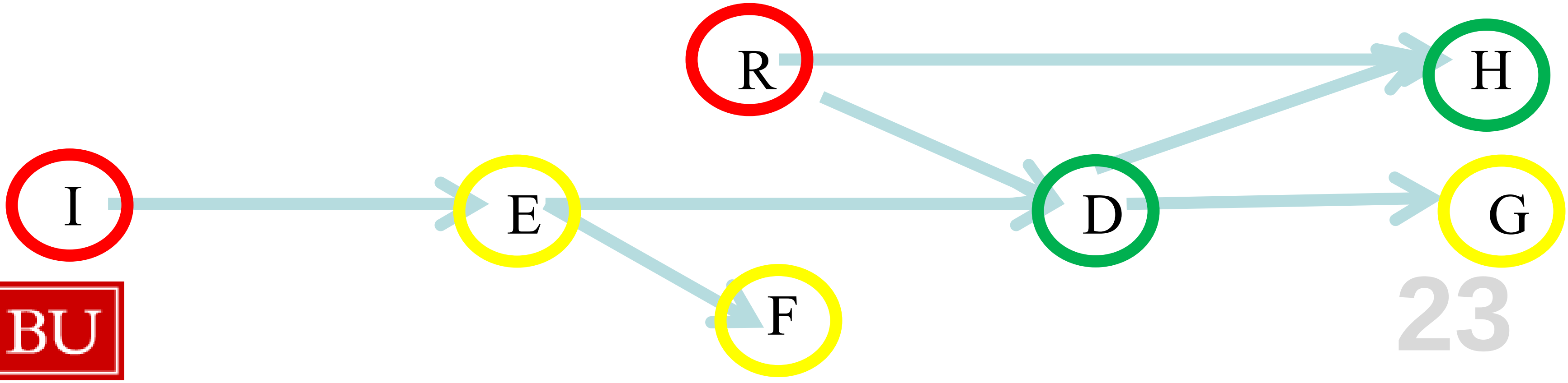
Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	8.2128	7.2595	9.2913
	Logit	7.9946	7.0651	9.0463
Cohort (Col1 Risk)	Mantel-Haenszel	5.0930	4.6082	5.6289
	Logit	5.0643	4.5876	5.5906

BU

Simulating DAGs

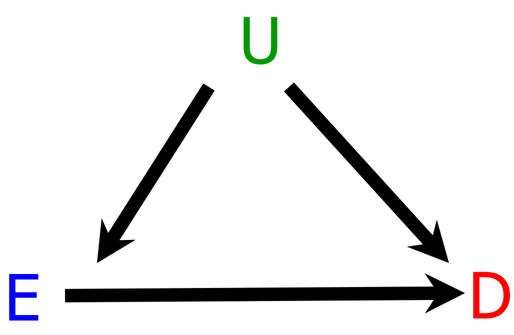
- Find the independent nodes and simulate
  - Specify probability
- Simulate nodes dependent on one arrow
  - Specify probability in all levels of the arrows the leads into the node
- Simulate nodes dependent on only two arrows, etc.
  - Specify probability in all levels of arrows that lead into the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



BU

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
  - RR_{UD} = 2.5, Pr(U+|E+) = 10% Pr(U+|E-) = 20%



Crude			U+			U-		
	E+	E-		E+	E-		E+	E-
D+	a 45	b 70	D+	A1	B1	D+	A0	B0
D-	c 255	d 630	D-	C1	D1	D-	C0	D0
Total	m 300	n 700	Total	M1 30	N1 140	Total	M0 270	N0 560

Unmeasured Confounders

- $RR_{CD} = 2.5$ and

$$A_1 = \frac{RR_{CD} M_1 a}{RR_{CD} M_1 + m - M_1} \qquad A_1 = \frac{2.5*30*45}{2.5*30+300-30}$$

$$B_1 = \frac{RR_{CD} N_1 b}{RR_{CD} N_1 + n - N_1} \qquad B_1 = \frac{2.5*140*70}{2.5*140+700-140}$$

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

Crude			**U+**			**U-**		
	E+	E-		E+	E-		E+	E-
D+	a 45	b 70	D+	A1 9.8	B1 26.9	D+	A0 35.2	B0 43.1
D-	c 255	d 630	D-	C1 20.2	D1 113.1	D-	C0 234.8	D0 526.9
Total	m 300	n 700	Total	M1 30	N1 140	Total	M0 270	N0 560

BU

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+) = 9.8/45, Pr(U+|E+,D-) = 20.2/255
  - Pr(U+|E-,D+) = 26.9/70, Pr(U+|E-,D-) = 113.1/630
- Code:
  - if E=**1** and D=**1** then U = rand('bernoulli', **9.8/45**);
  - else if E=**1** and D=**0** then U = rand('bernoulli', **20.2/255**);
  - else if E=**0** and D=**1** then U = rand('bernoulli', **26.9/70**);
  - else if E=**0** and D=**0** then U = rand('bernoulli', **113.1/630**);

Crude			**U+**			**U-**		
	E+	E-		E+	E-		E+	E-
D+	a 45	b 70	D+	A1 9.8	B1 26.9	D+	A0 35.2	B0 43.1
D-	c 255	d 630	D-	C1 20.2	D1 113.1	D-	C0 234.8	D0 526.9
Total	m 300	n 700	Total	M1 30	N1 140	Total	M0 270	N0 560

BU

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

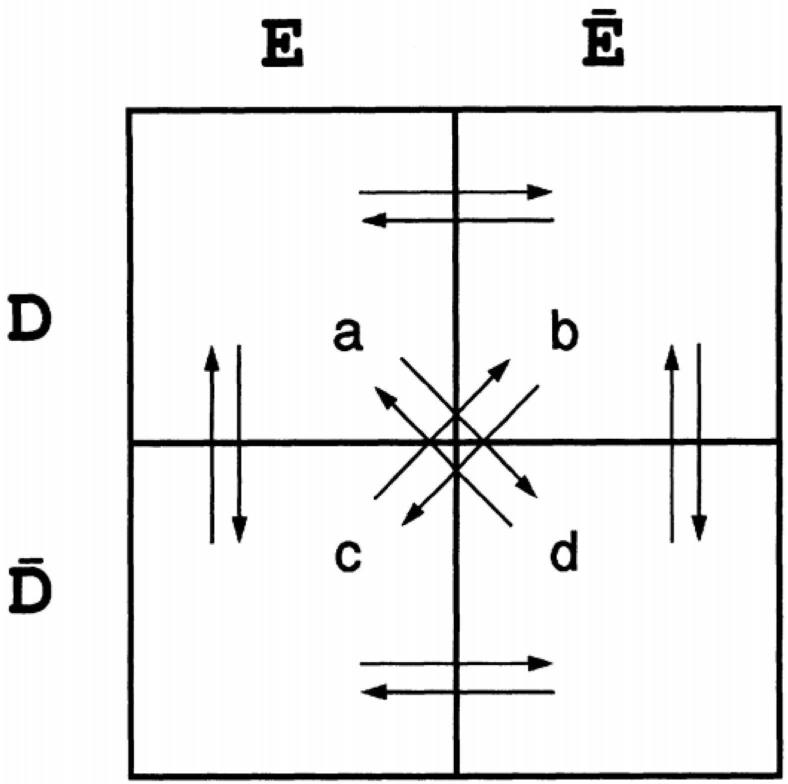


FIGURE 1. Rearrangement resulting from error in classification of exposure (E, $\bar{E}$) and outcome (D, $\bar{D}$).