Teaching Causal and Statistical Reasoning

Richard Scheines

and *many* others

Philosophy, Carnegie Mellon University
Outline

Today

1. The Curriculum
2. The Online Course
   - Modules
   - Case Studies
   - Support Materials
   - Causality Lab
3. Learning Studies

Tomorrow

1. The Causality Lab
   - Doing Exercises
   - Authoring Exercises
2. Pilot Studies
Motivation

• In College Curriculum:
  Statistical Methods ubiquitous – Causation rare

• Empirical Research Methods experience

• We have a theory to sell

• We got a grant – ugh.
Educational Goals

• Providing an extensive introductory treatment of the modern theory of causation and its relationship to statistical ideas

• Equipping students with the analytical tools needed to critically assess social and behavioral “studies” reported in the press

• Providing a foundation for more advanced work in Causal Bayes Networks
Causal and Statistical Reasoning Curriculum

1. Causation

2. Association and Independence

3. Causation $\rightarrow$ Association

4. Association $\rightarrow$ Causation
Causation

- **Foundations** (Events, Kinds of Events, Variables, Populations and Samples)

- **Causation Among Variables**
  - Deterministic Causation
  - Indeterministic Causation

- **Representation:**
  - Causal Graphs
  - Modeling Ideal Interventions
Direct Causation

X is a direct cause of Y relative to S, iff

\[ \exists z, x_1 \neq x_2 \quad P(Y \mid X \text{ set}= x_1, Z \text{ set}= z) \neq P(Y \mid X \text{ set}= x_2, Z \text{ set}= z) \]

where \( Z = S - \{X, Y\} \)

\[ X \rightarrow Y \]
Causal Graphs

Causal Graph $G = \{V, E\}$
Each edge $X \rightarrow Y$ represents a direct causal claim:
$X$ is a direct cause of $Y$ relative to $V$

Chicken Pox

Exposure $\rightarrow$ Rash

Exposure $\rightarrow$ Infection $\rightarrow$ Rash
Modeling Ideal Interventions

Interventions on the Effect

Post experimental System

Wearing Sweater  \[\rightarrow\]
Room Temperature  \[\times\]
Modeling Ideal Interventions

Interventions on the Cause

Post-experimental System

Wearing Sweater  Room Temperature
Interventions & Causal Graphs

- Model an ideal intervention by adding an “intervention” variable outside the original system

- Erase all arrows pointing into the variable intervened upon

---

Pre-intervention graph

Intervene to change Inf

Post-intervention graph?
Association and Independence

- Relative Frequency
- Conditional Relative Frequency
- Independence
- Conditional Independence
Causation → Association

D-separation

Equivalence

X1 → X2 → X3

X1 ← X2 ← X3

X2 → X3 → X1

X2 ← X3 ← X1

D-separation

X1 _||_ X3 | X2

X1 _||_ X3
Association $\rightarrow$ Causation

Problems for Causal Discovery:
- Underdetermination
- Confounding
- Measurement Error
- Sampling Variability (Statistics!)

Strategies for Causal Discovery
- Experiments (Interventions)
- Statistical Control (multiple regression, etc.)
- Search
Causal and Statistical Reasoning Online
www.phil.cmu.edu/projects/csr

Open Learning Initiative
http://oli.web.cmu.edu

• 16 Content Modules (full semester course)
• > 100 “Case Studies”
• Causality Lab
• Support Materials
  – Recitation Lessons
  – Tests
CSR

www.phil.cmu.edu/projects/csr

Demo
The Online Course

**Spring 2005 - CSR Demonstration Course**

Instructors: OLH HELP (oli-helplists.andrew.cmu.edu), Richard Scheines (scheines@andrew.cmu.edu)

Term: Spring 2005

Course Admit Code: cs1105

Before you begin, Test and Configure your system for use with this course.

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UNIT 1: CAUSATION</strong></td>
<td></td>
</tr>
<tr>
<td>Module 1: Causation; Preliminaries</td>
<td></td>
</tr>
<tr>
<td>Module 2: Causation Among Variables</td>
<td></td>
</tr>
<tr>
<td>Module 3: Indeterministic Causation</td>
<td></td>
</tr>
<tr>
<td>Module 4: Causal Graphs</td>
<td></td>
</tr>
<tr>
<td>Module 5: Interventions</td>
<td></td>
</tr>
<tr>
<td><strong>UNIT 2: ASSOCIATION AND INDEPENDENCE</strong></td>
<td></td>
</tr>
<tr>
<td>Module 6: Relative Frequency</td>
<td></td>
</tr>
<tr>
<td>Module 7: Conditional Relative Frequency</td>
<td></td>
</tr>
<tr>
<td>Module 8: Independence and Association</td>
<td></td>
</tr>
</tbody>
</table>
The Online Course: Roster and Gradebook

Spring 2005 - CSR Demonstration Course

Instructors: OLI HELP (oli-help@lists.andrew.cmu.edu), Richard Scheines (scheines@cmu.edu)
Term: Spring 2005
Course Admit Code: csrdemo05

5 registered students, 0 pending students, 1 suspended student

Registered Students

Registered students may fully participate in the course.

<table>
<thead>
<tr>
<th>Name</th>
<th>ID</th>
<th>E-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easterday, Matthew</td>
<td>matteasterday</td>
<td><a href="mailto:mwe@andrew.cmu.edu">mwe@andrew.cmu.edu</a></td>
</tr>
<tr>
<td>Gross, Justin</td>
<td><a href="mailto:jjgross@ANDREW.CMU.EDU">jjgross@ANDREW.CMU.EDU</a></td>
<td><a href="mailto:jjgross@ANDREW.CMU.EDU">jjgross@ANDREW.CMU.EDU</a></td>
</tr>
<tr>
<td>Rinderle, John</td>
<td>jar2</td>
<td><a href="mailto:jar2@andrew.cmu.edu">jar2@andrew.cmu.edu</a></td>
</tr>
<tr>
<td>Student, Demo</td>
<td>demostudent</td>
<td><a href="mailto:scheines@cmu.edu">scheines@cmu.edu</a></td>
</tr>
<tr>
<td>Tang, Adrian</td>
<td>bct</td>
<td><a href="mailto:bct@andrew.cmu.edu">bct@andrew.cmu.edu</a></td>
</tr>
</tbody>
</table>
The Online Course: Content Modules

Presenting Concepts

Now suppose we conduct an experiment in which we ideally intervene on X. Here are the seven post-manipulated graphs.

<table>
<thead>
<tr>
<th>POST-MANIPULATION GRAPHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRAPH 1</td>
</tr>
<tr>
<td><img src="image1" alt="Graph 1" /></td>
</tr>
<tr>
<td>X → Y</td>
</tr>
<tr>
<td>INTERVENTION</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GRAPH 4</th>
<th>GRAPH 5</th>
<th>GRAPH 6</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image4" alt="Graph 4" /></td>
<td><img src="image5" alt="Graph 5" /></td>
<td><img src="image6" alt="Graph 6" /></td>
</tr>
<tr>
<td>X → Y</td>
<td>X → Y</td>
<td>X → Y</td>
</tr>
<tr>
<td>INTERVENTION</td>
<td>INTERVENTION</td>
<td>INTERVENTION</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GRAPH 7</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7" alt="Graph 7" /></td>
</tr>
<tr>
<td>X → Y</td>
</tr>
<tr>
<td>INTERVENTION</td>
</tr>
</tbody>
</table>

Comprehension Checks

The General Theory 1

[Question 1] [Question 2] [Question 3] [Question 4]

Attempt 1 for this question

**Question 1**

In which of these post-manipulation graphs is there a causal connection between X and Y?

- [ ] A.
  ![Graph A](image8)
  INTERVENTION
  ![Node X](image9) ![Node Y](image10)

- [ ] B.
  ![Graph B](image11)
  INTERVENTION
  ![Node X](image12) ![Node Y](image13)

So the ideal intervention on X eliminates all causal connections between X and Y except for one: a causal path from X to Y.
Asthma 'linked to obesity'

Date: April 27, 1999

Source: http://news.bbc.co.uk/hi/english/health/default.stm

Copyright: 1999 BBC

Concepts
- variables
- causal graphs
- confounders

Keywords
- asthma
- obesity

Researchers assessed data from the 1970 British Cohort Study, an ongoing study of almost 9,000 people born between 5 and 11 April in 1970 whose health and behaviour have so far been followed up at the ages of 5, 10, 16 and 26 years.

They found that the fatter the adult, the greater the likelihood of asthma.
The Causality Lab

www.phil.cmu.edu/projects/causality-lab
Support Materials
Recitation Lessons, Tests
CSR Usage

- ~ 2,800 total students
- ~ 75 different courses
- ~ 45 Institutions

Disciplines:
- Philosophy
- Statistics
- Psychology
- Political Science
- Math
- Management
- Nursing
- Speech
- Economics
- Marketing
CSR Evaluation

- How do students fare with online vs. lecture delivery of identical material?
- What factors affect the pedagogical outcome?
  e.g., face-to-face attendance, time online, exercises attempted, etc.
- What does it cost?
Experiments

2000: Online vs. Lecture, UCSD
- Winter (N = 180)
- Spring (N = 120)

2001: Online vs. Lecture, Pitt & UCSD
- UCSD - Winter (N = 190)
- Pitt (N = 80)
- UCSD - Spring (N = 110)
Online vs. Lecture Delivery

• Online:
  – *No* lecture / one recitation per week
  – Required to finish approximately 2 online modules / week

• Lecture:
  – 2 Lectures / one recitation per week
  – Printed out modules as reading – extra assignments

• Same Material, same Exams:
  – 2 Paper and Pencil Midterms
  – 1 Paper and Pencil Final Exam
Experimental Design

Registered Students

A) Online

Stratified Random Draw

B) Lecture (Control)

1st Day Pre-test, Choice

C) Lecture

Lecture

Online

A vs. B -- Main effect

B vs. C -- Selection Bias
• Online students averaged 1/2 a Stdev better than lecture students (p = .059)
• Factors affecting performance: Practice Questions Attempted
• Cost: Online costs 1/3 less per student
Recitation attendance is more important than lecture attendance
Printing and Voluntary Comprehension Checks: 2002 --> 2004

2002

pre -> print
pre -> final
print -> voluntary questions
final -> voluntary questions
voluntary questions -> quiz

2004

pre -> print
pre -> final
print -> voluntary questions
final -> voluntary questions
voluntary questions -> quiz
Student Behavior Patterns Relevant to Learning

Time on task:
- time *engaged*
- pattern of work (deadline proximate work only, etc.)

Activities:
- Reading
- Comprehension checks
- Simulations
- Lab exercises

Work Patterns, Help seeking, etc.
Time Spent on Learning Pages

Fall 04:
• 2 classes, 40 Students,
• > 8,000 learning page visits

Filtered (to 6,418):
• 5 seconds < page visits < 600 seconds
• Only pages hit at least 5 times
• Only sessions including at least 5 page visits
• Only students with at least 5 sessions

Variables:
• $Page_{demand}(i)$: mean time spent over all visits to learning page $i$
• $Session(j)$: mean time spent on learning pages during session $j$
• $Student(k)$: mean time spent on learning pages by student $k$

Visit_length on page $i$ during session $j$ by student $k = f(Page_{demand}(i), Session(j), Student(k), \epsilon)$
Time Spent on Page

Visit_length on page i during session j by student k =
\[ f(\text{Page\_demand}(i), \text{Session}(j), \text{Student}(k), \varepsilon) \]

Linear Regression:

Visit_length = .838 Session + .837 Page\_demand + .141 Student

R-square = .315
R-square (w/o Student) = .314
References

Causal and Statistical Reasoning Online
www.phil.cmu.edu/projects/csr

Open Learning Initiative
http://oli.web.cmu.edu
