2014 Schield eCOTS 1

TWO BIG IDEAS FOR TEACHING BIG DATA

Coincidence & Confounding

by
Milo Schield
Augsburg College, USA
Electronic Conference on Teaching Statistics (E-COTS)
May 20, 2014.
www.StatLit.org/pdf/2014-Schield-eCOTS-Slides.pdf

2014 Schield eCOTS 2

Start up

How many participants are online? __________

Q1: When teaching introductory statistics, who chooses your text? ___ Teacher ____ Teachers together ___ Others

Q2. What fraction of a one-semester introductory statistics course should focus on *coincidence* and *confounding*?
____ 0 - 5%; _____ 5 -15%; _____ 15 - 30%
____ 30 - 50% _____ At least half

3

Big Data and Big Ideas

In big data,
1. *Coincidence* is a much bigger problem.
2. *Confounding* is often the #1 problem.

2014 Schield eCOTS 4

True Confession

I have been teaching introductory statistics for over two decades. I have a confession.

2014 Schield eCOTS 5

Survey Question 3

How many introductory statistics textbooks use *coincidence* or *chance* to support the claim that *association is not causation*?

<u>Response</u> <u>Choice</u>
_____ None
_____ One or two
_____ Three-to-six
_____ More than half a dozen.

6

2014 Schield eCOTS 7

Demonstrating Coincidence

Seems impossible!

Three cases:

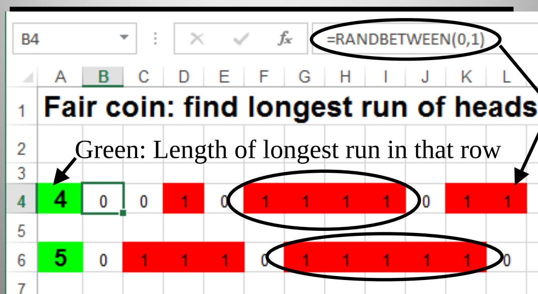
1. Run of heads
2. Grains of Rice
3. Birthday Problem

2014 Schield eCOTS 8

#1 Run of Heads

A common occurrence. Given enough tries, the unlikely is expected.

Flip coins in rows. 1=Heads (Red fill) Adjacent Red cells is a Run of heads.



Source: www.statlit.org/Excel/2012Schield-Runs.xls

Run of 4 heads: 1 chance in 2^4 = 1/16 Run of 19 heads: 1 in 2^19 = 1/524,288

Source: www.statlit.org/Excel/2012Schield-Runs.xls

11

Consider a run of 10 heads? What is the chance of that?

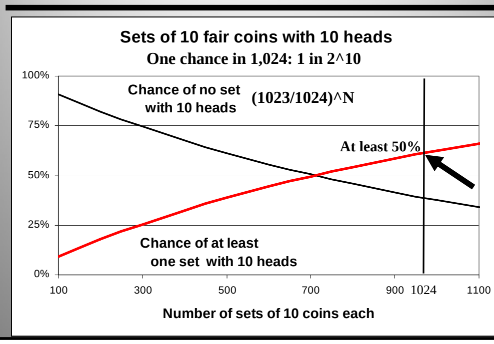
Question is ambiguous! Doesn't state context!

1. Chance of 10 heads on **the next 10 flips**?
   p = 1/2; k = 10.
   P = p^k = (1/2)^10 = one chance in 1,024
2. What is the chance of at least one set of 10 heads [**somewhere**] when flipping 1,024 sets of 10 coins each? At least 50%.*

* Schield (2012)

12

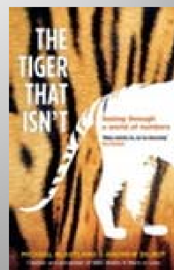
Coincidence increases as data size increases



2014 Schield eCOTS

#2 Grains of Rice
Blastland: *The Tiger That Isn't*

With rice scattered in two dimensions, people can often see memorable shapes.



After this webinar, check out this Excel scattered-rice demo with 1 chance in 100 per cell:

www.StatLit.org/Excel/2012Schield-Rice.xls

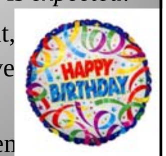
2014 Schield eCOTS 14

#3: The "Birthday" Problem: Chance of a matching birthday

Richard von Mises (1938)

In a group of 28 people,
a birthday match is *expected*.

The trick is to show it,
– not just to prove



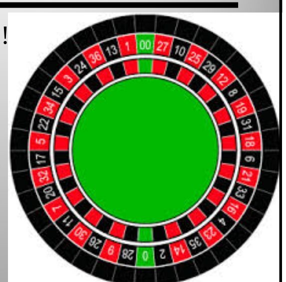
Try this Excel den

www.StatLit.org/Excel/2012Schield-Bday.xls

15

Law of Very-Large Numbers

Not Law of Large Numbers!



Qualitative form:
The unlikely is almost certain given enough tries.

Quantitative form:
Event: one chance in N.
In N tries, one event is 'expected' and is more likely than not. Schield (2012)

2014 Schield eCOTS

Coincidence Outcomes

Students must "see" that coincidence
- may be more common than expected
- depends on the context
- may be totally spurious
- may be a sign of causation

2014 Schield eCOTS 17

Survey Question 4

Would you teach *coincidence* in an introductory statistics class?

Response	Choice
_____	No
_____	Possibly
_____	Probably
_____	Almost certainly

18

Second Big Idea: Confounding

As sample size increases,
- Margin of error decreases,
- Coincidence increases (becomes more likely)
- Confounding remains unchanged.

Big data doesn't minimize confounding.
If anything, Big Data gives unjustified support for confounder-spurious associations.

19

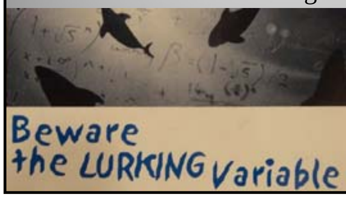
Second Big Idea: Confounding

CLAIM

Simpson's paradox (sign reversal or confounding)

•is incidental when modelling or forecasting,

•dominates when searching for causes.



20

Modeling NAEP data

Based on 2001 NAEP Math 4 Scores

	Low$ (0)	High$ (1)	Total
Utah (0)	209	234	228
Okla (1)	218	244	224
Total	214 →	239	226

$ indicates student has low or high family income

Source: www.StatLit.org/pdf/2004TerwilligerSchieldAERA.pdf
Data at www.StatLit.org/Excel/2014-Schield-eCOTS-Data.xls

21

Forecast with Confounder; Reversal is Incidental

Data based on 2001 NAEP 4th Grade Math Scores.
Compare Utah (0) and Oklahoma (1)

Score = 228 - 4.5*State			Score = 208.7 + 9.5*State + 25.0*Income	
Regression Statistics				
R Square	0.02	→ Increase →	R Square	0.42
Standard Error	16.23	Decrease	Standard Error	12.48
p-value (Intercept	0.00	Decrease	Pvalue (Intercept)	0.00
p-value (STATE)	0.02	Decrease	p-value (STATE)	0.00
Observations	300		p-value (INCOME)	0.00

Adding more factors typically improves the quality of the model

22

Explain with Confounder; Reversal is Essential

Based on 2001 NAEP 4th Grade Math Scores

	Low$ (0)	High$ (1)	Total	%High$
Utah (0)	209 ↓	234 ↓	228 ↑	78%
Okla (1)	218	244	224	22%

Causal Question:
Which State has the better education system?

Score = 228.3 - 4.5*State → Utah (0) is better

Score = 208.7 + 9.5*State + 25.0*Income → Oklahoma (1) is better

Data at www.StatLit.org/Excel/2014-Schield-eCOTS-Data.xls

23

Teaching Confounding: Two Big Reasons Not To…

(1) Disrespect

(2) Prerequisites

An open mind is the prerequisite to gaining knowledge.

24

Teaching Confounding: Reasons To…

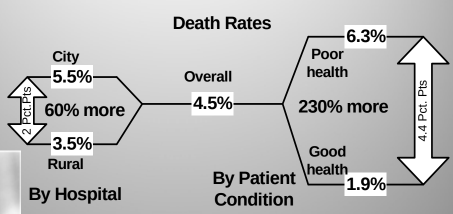
#1: The Cornfield conditions[1] set a minimum on the size confounder that can negate or reverse an association. Schield (1999). These conditions can offset excessive skepticism/cynicism.

#2: When the predictor and confounder are binary, there are graphical techniques[2] that allow students to work problems without software and without a second course in regression. Schield (2006)

This material has been taught for over 10 years.

25

#1: Using Cornfield's Condition



Cornfield's condition: To reverse an association, the confounder must be *bigger* than the association.

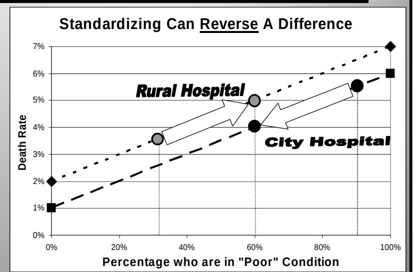
26

#2: Standardizing with binary predictor and confounder

For a step-by-step overview of this new graphical standardizing procedure:

•See Schield (2006).

•Listen to audio; view the slides.



Audio: www.statlit.org/Audio/2009StatLitText-Overview-Ch3.mp3
Slides: *www.statlit.org/pdf/2009StatLitTextHandoutCh3.pdf*

27

Conclusion

Many – if not most – big-data users want causal explanations (C.f., business intelligence). Modeling and prediction are just a means to this end.

To be relevant for these users of Big Data,

1. We must focus more on Coincidence & Confounding. These are two big influences on many statistics. Our students deserve a broader education.

2. We must say more about causes than "Association is not Causation." We must introduce confounding, the Cornfield conditions and standardization.

28

Questions 5 and 6

Q5. Business majors deal with causes. What fraction of a *Business Statistics* course should focus on *coincidence* and *confounding*?

____ 0 - 5%; _____ 5 -15%; _____ 15 - 30%
____ 30 - 50% _____ At least half

Q6. If you taught *Business Statistics,* would you investigate an introductory textbook with a strong emphasis on *coincidence* and *confounding*?

_____ No way; ______ Conceivably; ____ Possibly;
_____ Probably; _____ Almost certain.

29

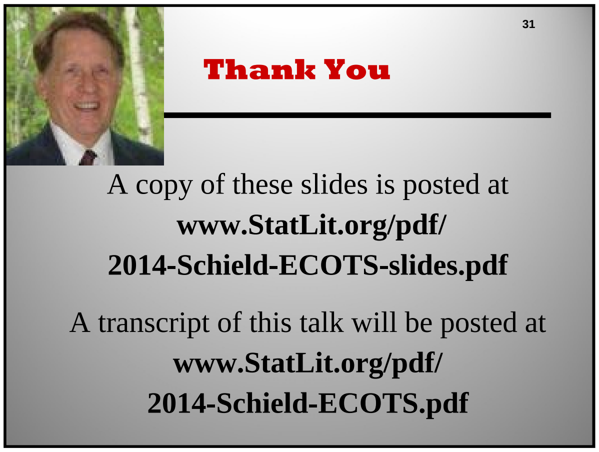
References

1. Schield (1999). Simpson's Paradox and Cornfield's Conditions, ASA Proceedings Statistical Education. www.StatLit.org/pdf/1999SchieldASA.pdf.
2. Schield (2006). Presenting Confounding Graphically Using Standardization. *STATS magazine*. www.statlit.org/pdf/2006SchieldSTATS.pdf
3. Schield (2012). Coincidence in Runs and Clusters www.statlit.org/pdf/2012Schield-MAA.pdf
4. Terwilliger and Schield (2004). Frequency of Simpson's Paradox in NAEP Data. AERA. See www.StatLit.org/pdf/2004TerwilligerSchieldAERA.pdf

30

Suggested Readings

1. Pearl, Judea (2000). Simpson's Paradox: An Anatomy. http://bayes.cs.ucla.edu/R264.pdf
2. Pearl, Judea (2014). Understanding Simpson's Paradox. *The American Statistician*, 2/2014, V68, N1 http://ftp.cs.ucla.edu/pub/stat_ser/r414-reprint.pdf
3. Pearl, J. (2014). Statistics and Causality: Separated to Reunite. Commentary. Health Service Research. http://ftp.cs.ucla.edu/pub/stat_ser/r373-reprint.pdf
4. Gelman blog (2014). On Simpson's Paradox. http://andrewgelman.com/2014/02/09/keli-liu-xiao-li-meng-simpsons-paradox/


31
Thank You
A copy of these slides is posted at
www.StatLit.org/pdf/
2014-Schield-ECOTS-slides.pdf
A transcript of this talk will be posted at
www.StatLit.org/pdf/
2014-Schield-ECOTS.pdf