

Welcome to this ECOTS invited webinar: Two big ideas for teaching big-data.

Thanks to Jean and Michelle for heroic efforts in trying to make the originally-scheduled webinar work despite the technical difficulties. This is a redo without any audience.

In that aborted session we started with 54 participants and peaked at 67 before stopping since no one could hear my voice. We did get responses to three survey questions.

Question 1: “When teaching introductory statistics, who chooses your textbook?” Possible answers with percentages: The teacher [54%]; Teachers together [42%]; Someone else [29%]. Obviously these choices weren’t exclusive.

Question 2: “What fraction of a one-semester introductory statistics course should focus on coincidence and confounding?” 0-5% [31%]; 5-15% [44%]; 15-30% [19%]; 35-50% [nil]; more than 50% [6%].

I didn’t expect that a fourth of those responding would recommend spending at least 15% of the course (almost a fifth of the course) on coincidence and confounding. I’d probably answer in the 0 to 5% category. I just don’t have time to put another topic – or two topics -- into my overloaded intro statistics course. But the allocation certainly depends on our priorities.

I am here to talk about making coincidence and confounding a priority for your intro statistics course and for big data.

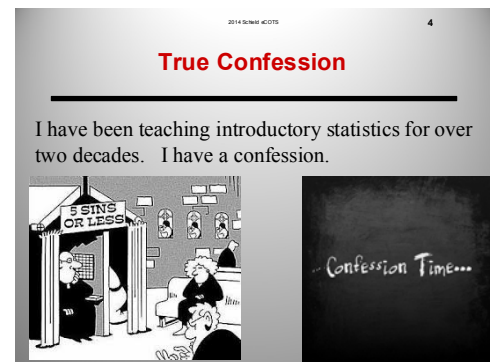
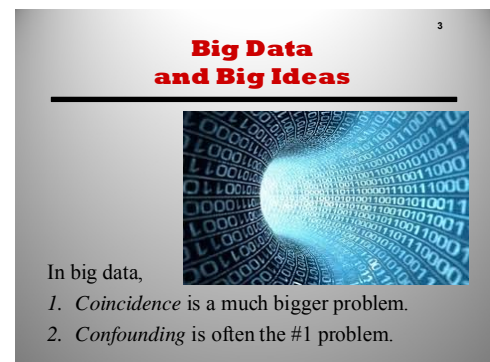
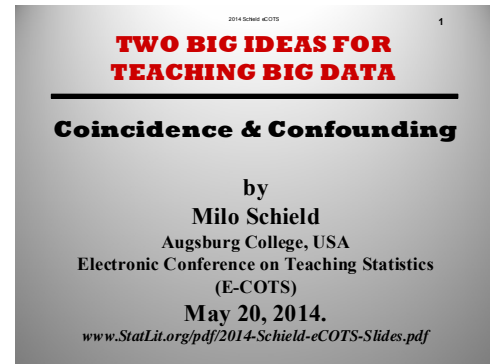
Slide 3: For me “big-data” is any data set in which all the associations are statistically significant. This means that statistical significance doesn’t have much value with big-data.

But, in big data coincidence and confounding are typically the big statistical problems.

Slide 4: I have a true confession. I’ve been teaching statistics a long time. I always note that “association is not causation.” My examples always involved confounding: the Berkley sex discrimination case; ice-cream and burglaries in the summer. But my course never involved confounding. It was all about randomness: sampling error, margin of error and confidence intervals. In a sense, I was guilty of bait and switch.

I introduced my students to one kind of problem and then I switched to solving a different kind of problem. I want to show you how I start my course now using coincidence. But first another poll

Question 3: How many intro statistics textbooks use coincidence or chance to support the claim that “association is not causation”? Answers: None [6%]; One or two [41%]; Three to six [24%]; More than six [29%]. My answer was “None”, but only 6% agreed with me. Either I’m not well-read, or I’ve written



the question wrong or “something”. Send an e-mail to Schield@Augsburg.edu with the names of textbooks. I’d certainly appreciate getting the education.



Slide 6: Here are some humorous examples of coincidence. For me, coincidence is an unlikely combination of events that is memorable. When you watched the royal wedding, did you notice the many similarities shown in these pictures? Now, I got these pictures off the web, so they may be photo-shopped for all I know. Still, coincidences are memorable; we all have stories of amazing coincidences.

Slide 7: How are we going to demonstrate a coincidence in our classroom – on demand – every time? That seems impossible! I’m saying it isn’t impossible. I’m going to introduce three examples of coincidence that I use to start my intro statistics course. In each example the underlying probabilities are known. The three spreadsheets involve Run of heads, Grains of Rice and the Birthday Problem. Let’s start with a simple case of coincidence: a run of heads in flipping fair coins.

Demonstrating Coincidence

Seems impossible!

Three cases:

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

Slide 8: We all know about a run of heads. Students don’t expect long runs; they think long runs are a sign of non-randomness. Auditors use the absence of long runs as a sign of fraud.



Slide 9. Examine this Excel worksheet that I created for my students. It’s free. Let your students use it. Have fun!

I used the RandBetween() function to generate random outcomes. Each cell in a row is either 1 or 0: heads or tails. Cells with heads are conditionally formatted in red. A run is a group of red cells that are touch.

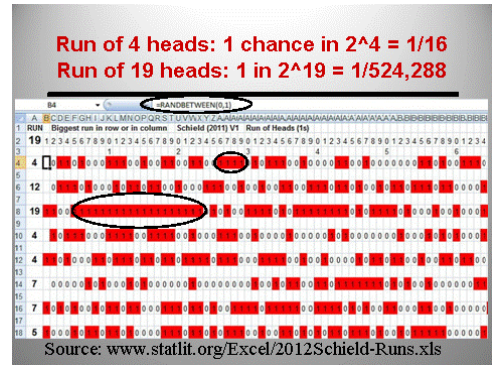
Do you see the number in green at the left end of each row? That is the length of the *longest run* in that row. In row 4, the longest run is length four. In row 6, the longest run is length five. Now that you understand how the spreadsheet was designed, let’s take a look at the full spreadsheet.

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

	A	B	C	D	E	F	G	H	I	J	K	L
1	Fair coin: find longest run of heads											
2	Green: Length of longest run in that row											
3												
4	4	0	0	1	1	1	1	0	1	1		
5												
6	5	0	1	1	1	1	1	1	0			
7												

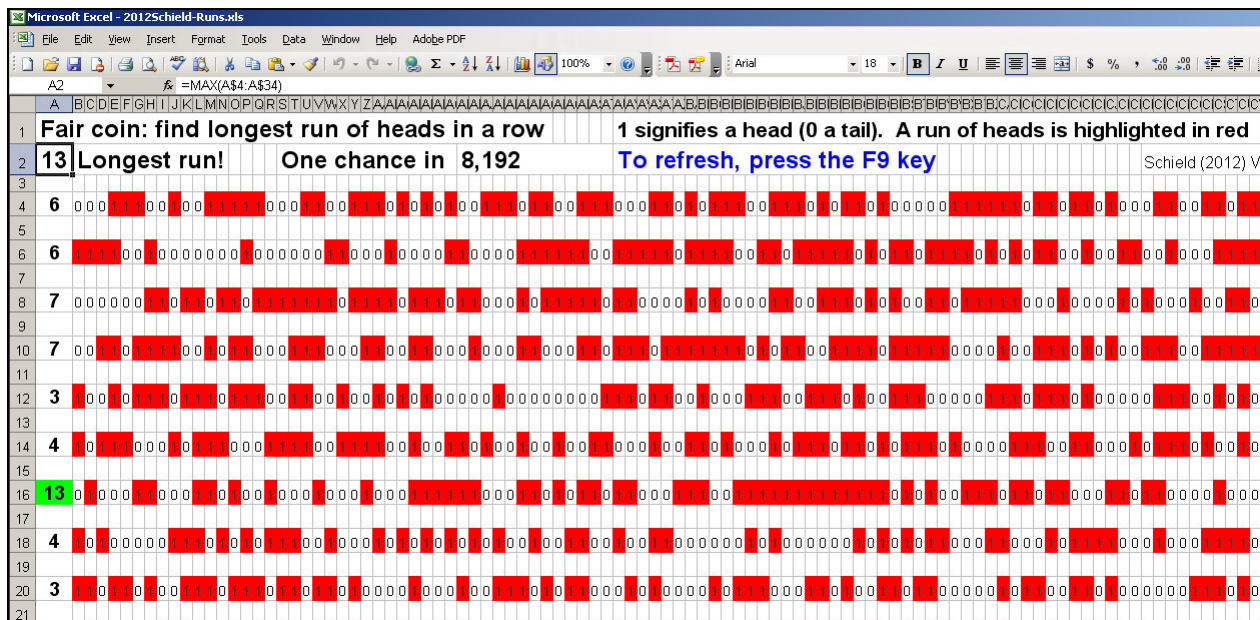
Source: www.statlit.org/Excel/2012Schield-Runs.xls

Slide 10: The number in the upper left hand corner is the longest run on the entire worksheet. In the first row, the longest run is length 4: one chance in 16. In the second row, the longest run is length 12: one chance in 2^{12} .

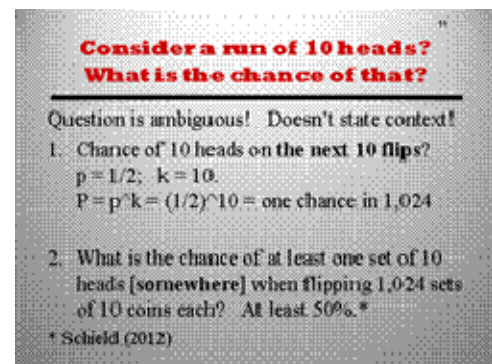


In the third row, the longest run is length 19: one chance in over 500,000. We pressed F9; we got a run of 19 heads. We just had a 500,000 year flood this year. The students enjoy pressing F9. As noted on this slide, this spreadsheet is available at www.StatLit.org/Excel/2012Schield-Runs.xls

Now open the Runs spreadsheet: Enable Editing. Refresh by pressing the F9 key. Students love playing with this spreadsheet. They are dimly aware that these long runs are amazing coincidences – and yet they are happening every time. Hold that thought.



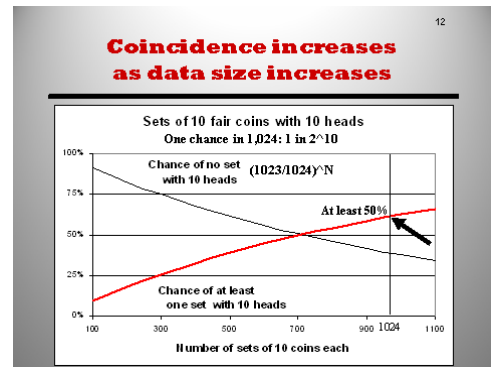
Slide 11. Suppose we point to a run of 10 heads and ask students “What is the chance of that?” At this point my students are confused. They know these long runs are extremely unlikely. But they also realize these long runs are happening most of the time.



Students don’t realize the question is ambiguous. It doesn’t state the context. If we are asking the chance of 10 heads on the next 10 flips (or at a pre-specified starting point), the answer is one chance in 1,024.

But if we are asking the chance of a set of 10 heads somewhere when you flip 1024 sets with 10 heads each, you can expect at least one. And it can be shown that the chance of getting at least one set of 10 heads is at least 50%: it’s “more likely than not”. See Schield (2012) for details. My students can remember this.

Slide 12 shows that as the sample size increases, the chance of no set of 10 heads decreases and the chance of at least one set of 10 heads increases. As statisticians, we are usually interested in the point where the lines intersect at 50%.



My students can't deal with that level of math. But, they can remember what is expected when $N = 1/p$.

And with $N = 1/p$, getting at least one set of 10 heads is more likely than not. See Schield (2012).

Slide 13: A second example of coincidence was proposed by Michael Blastland in his great book, *the Tiger That Isn't*. If you haven't read this book, I strongly recommend it. Michael noted that people see patterns in a random distribution of rice grains on the floor.

#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

I simulated this random distribution of rice grains in an Excel spreadsheet: www.StatLit.org/Excel/2012Schield-Rice.xls

Let your students play with this spreadsheet. Each cell has one chance in 10 of coming up with a number between zero and 9. If the random number is a nine that is considered to be a grain of rice and that cell is conditionally formatted in red. Here we have two red cells touching. What is the chance of that? One chance in 100. Here we have three red cells touching. What is the chance of that? One chance in a thousand. Here we have four red cells touching; one chance in 100,000. That's incredible. Students love pressing F9. They find these unlikely coincidences *every time*.

Rice-10 sheet

Find the largest group of Red cells that are touching each other.

a. touching on sides in a row

b. touching on top or bottom in a column

c. touching on tips or points in a diagonal

Conditional Format: Equal to 9, Fill Red, TextColor White.

Home tab: Cells/Format/Col-Width=2

Slide 14: My third example of coincidence is the birthday problem created by Richard von Mises in 1938. We know that with 23 people we can get a match at least half the time. But getting 23 involves more math than my students can follow.

I'm interested in using the expected value. The smallest group with more than 365 combinations involves 28 people. With 28 people a match of birthdays (month and day) is expected. But my students aren't generally convinced; they cannot see 365+ combinations.

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

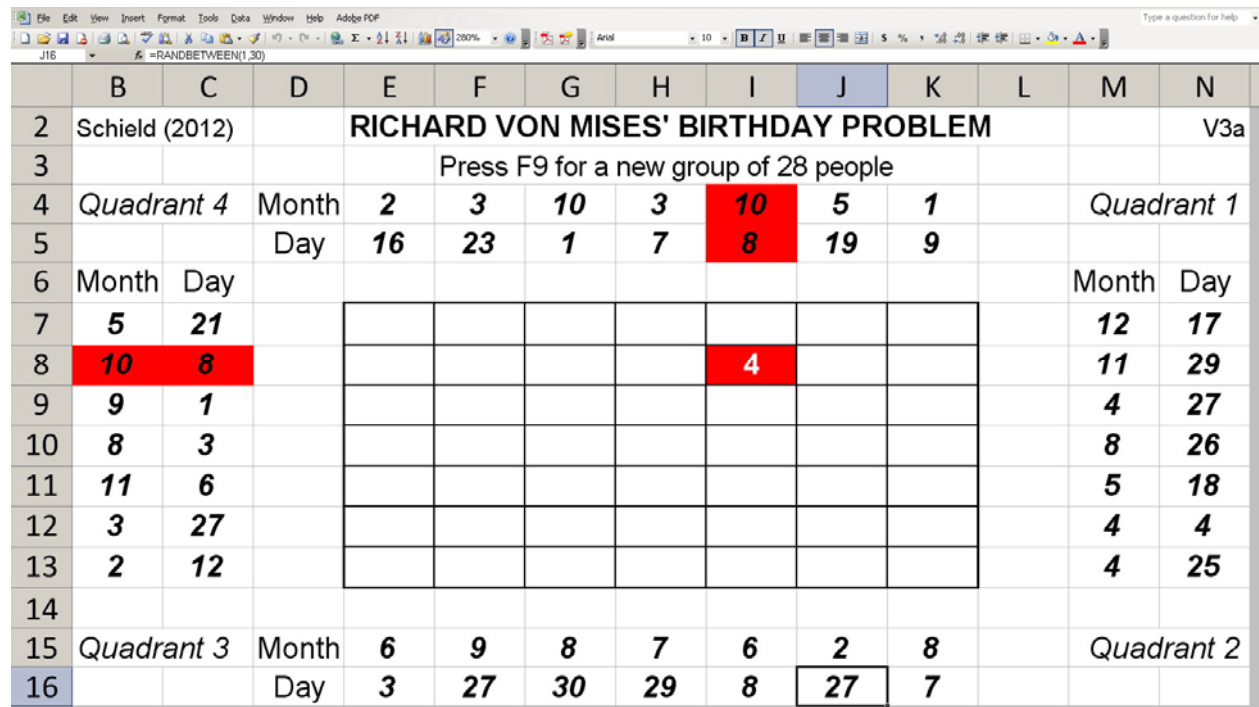


Richard von Mises (1938)
In a group of 28 people, a birthday match is *expected*.
The trick is to show it, — not just to prove



Try this Excel den
www.StatLit.org/Excel/2012Schield-Bday.xls

My students do understand that the chance of this event is one in 365. They understand that in 365 random pairs we can "expect" one match. All we need to do is to **show** students that there are more than 365 matches or pairs with 28 people. To show them, I created a spreadsheet which you can access on the web at www.StatLit.org/Excel/2012Schield-Bday.xls



	B	C	D	E	F	G	H	I	J	K	L	M	N
2	Schield (2012)		RICHARD VON MISES' BIRTHDAY PROBLEM										V3a
3	Press F9 for a new group of 28 people												
4	Quadrant 4		Month	2	3	10	3	10	5	1	Quadrant 1		
5			Day	16	23	1	7	8	19	9			
6	Month	Day										Month	Day
7	5	21										12	17
8	10	8										11	29
9	9	1										4	27
10	8	3										8	26
11	11	6										5	18
12	3	27										4	4
13	2	12										4	25
14													
15	Quadrant 3		Month	6	9	8	7	6	2	8	Quadrant 2		
16			Day	3	27	30	29	8	27	7			

There are seven people on each of the four sides of a table. Each person is identified with a random birthday: a random month and a random day. [All months are assumed to have 30 days. You can improve the spreadsheet if you want.]

In this spreadsheet there is a number in the central grid highlighted with a red fill: the number 4. Note that there are four quadrants: #1 in the upper right; #2 in the lower right, #3 in the lower left and #4 in the upper right. In this case, quadrant four has a match between the Oct 8 birthday in the same row on the left and the Oct 8 birthday in the same column at the top.

How many combinations are there in quadrant four between the 7 people on the top and the 7 on the left side? 49! There are 49 unique pairs for each of the four quadrants. That's almost 200 combinations for all four quadrants. Students can see this quickly.

The second kind of combination involves matches on opposite sides. Either a left-right match or else a top-bottom match. Again there are 49 unique combinations for each of the two opposite-side matches. With almost 100 new combinations, we are now up to nearly 300 combinations in total.

There third kind of combination involves matches on the same side. It's a little harder to see 21 combinations ($6+5+4+3+2+1$) for each of the four same-side matches. With around 80 new combinations, we now have about 380 unique pairs in total. The exact total is 378 pairs. Once students actually SEE more than 365 pairs with only 28 people, they are really convinced. Something with one chance in 365 is *expected!*

Slide 15: Suppose you show your students these three examples: run of heads, grains of rice and birthday problem. Does that mean your students will understand that some associations are not causation: that some associations could be spurious – just *coincidence*? My students don't. They just have three specific examples. They need a memorable principle to integrate these specifics together.

Slide 15: I introduce what I call *the Law of Very Large Numbers*. This law is not the same as the Law of Large Numbers. [This *Law of Very Large Numbers* parallels the *Law of Truly Large Numbers* introduced by Persi Diaconis and Frederick Mosteller.]

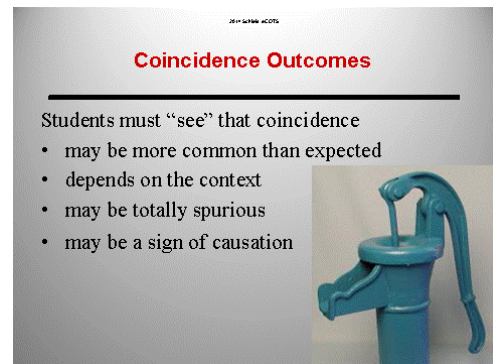
In its qualitative form, this Law of Very Large Numbers says that “The unlikely is almost certain given enough tries.” In its quantitative form, this law says “If you have an event with one chance in N and if you make N tries, then that event is expected AND it happens more than 50% of the time.” My students can understand this. It is almost intuitive: they can remember it and they can use it.

[The first part of this Law of Very Large Numbers is just a special case for the mean of the Binomial distribution: $E(k) = n \cdot p$. When $n = 1/p$, then $E(k) = (1/p) \cdot p$ which always equals one.

The second part of this law is extremely important. For a binomial distribution, $P(X=k) = \frac{n!}{k!(n-k)!} [p^k][(1-p)^{(n-k)}]$. $P(X=0) = \frac{n!}{n!} [p^0][(1-p)^n] = (1-p)^n$. When $n = 1/p$ or $p = 1/n$, then $P(X=0) = (1-1/n)^n$. For $n = 2$, $P(X=0) = (1/2)^2$ or 0.25. For $n = 10$, $P(X=0) = .9^{10} = 0.3487$. For $n = 100$, $P(X=0) = 0.99^{100} = 0.366$. Euler's constant $= e = \lim (1+1/n)^n$ as $n \rightarrow \infty$. Note that $1/e = 0.367879$. Schield (2012) argued that $P(X=0)$ approaches $1/e$ from below as n increases. Thus, the complement, $[1-P(X=0)]$, must approach $1-1/e$ from above. Since $1-1/e = 0.632$, $P(X>0)$ is always greater than 0.5 when $n = 1/p$.

Slide 16: What are the desired coincidence outcomes?
Students should see four things. (1) Coincidence may be more common than they realized. (2) Coincidence depends on context – before or after the fact. Did we set the target before we did the shooting or after? (3) Coincidence may be totally spurious – just random chance.

(4) Coincidence may still be a sign of causation. The pump handle reminds us of an important event in the history of statistics and statistical education. There is a pump in London on Broad Street (now Broadwick Street). This is near where in 1854 Dr. John Snow said the grouping of cholera deaths around a pump was not just coincidence – but was really a sign of causation. Snow's methodical analysis marked the beginning of epidemiology.



Slide 18: Now turn to our second big idea: *confounding*. To review: As sample size increases, margin of error decreases, and coincidence increases (it becomes more likely) but confounding remains unchanged. Suppose we noted that men were more likely to die of lung cancer than women. Getting more data might not change this comparison as long as the new data is the same kind as the prior. Getting more data wouldn't offset the influence of a relevant confounder: being a life-long smoker. If we never asked "Are you a life-long smoker?" and took it into account, we might mistakenly conclude that gender was largely responsible for this difference in lung cancer death rates.

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.

Big data doesn't minimize confounding. If anything, big data gives unjustified support for associations that may be spurious: chance spurious or confounder-spurious. [A bank analytics manager, Ryan Dunlop, noted that big data 'fishing expeditions' were often unproductive. A specific question was much more likely to generate useful results.]

Slide 19: The most interesting form of confounding is Simpson's Paradox. Simpson's Paradox is a sign reversal of a predictor after taking into account a relevant but extraneous factor. Here is my claim: In observational studies Simpson's Paradox reversal is incidental when modeling or forecasting. But such a reversal dominates when searching for causes.

19

Second Big Idea: Confounding

CLAIM

Simpson's paradox (sign reversal or confounding)

- is incidental when modelling or forecasting,
- dominates when searching for causes.

[Here is my argument. Including a confounder (a related factor) in a model *can* do two different things: (#1) improve the quality and predictions of a model; (#2) change the sign and size of other predictors. In modeling or forecasting, #1 (improving quality) dominates. In a causal analysis, #2 (changing the sign and size of predictors) dominates.]

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

To show this, I am using some data published by the National Assessment of Educational Progress (NAEP). This data is for US fourth graders taking a national math test in 2001.

Slide 20: The average score of the Utah students (228) was higher than that of the Oklahoma students (224). The data is broken out by student family income. The average score of students from high income families (239) was much higher than that from low-income families (214).

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		Standard Error	12.48
p-value (Intercept)	0.00	Decrease	p-value (Intercept)	0.00
p-value (STATE)	0.02		p-value (STATE)	0.00
Observations	300		p-value (INCOME)	0.00

Adding more factors typically improves the quality of the model

Slide 21: I generated detail data that matched these summary statistics. This data is at www.StatLit.org/Excel/2014-Schield-ECOTS-Data.xls. When we regress these NAEP math scores by state, we get the regression statistics shown on the left.

Adding more factors typically improves the quality of the

model. When we regress math scores by state and family income we get the regression statistics shown on the right. Compare the results. R-squared increased dramatically so the model fits better. Standard Error and p-value decrease, so predictions will be more accurate. The reversal in the sign of the coefficient for the STATE variable is incidental.

Slide 22: Now consider the same data analysis from a causal perspective. Which state has the better educational system? If we analyze just by State, Utah's schools seem better than Oklahoma's. But if we include student family income, Oklahoma's schools seem better than Utah's. Utah looked better originally because it had a much higher percentage of students from high-income families than did Oklahoma. This sign reversal for STATE is Simpson's Paradox. The significance of a sign reversal depends on what kind of analysis you do.

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/Exce1/2014-Schield-eCOTS-Data.xls

Slide 23: I claim we should teach more on confounding. But, I want to acknowledge there are two big reasons **NOT** to teach confounding in an introductory course.

#1: The first is disrespect. A colleague noted that students might have less respect for our discipline if they knew how easily statistics and statistical significance could be changed after taking into account an extraneous factor.

#2: We don't want to model bad practice. Regression involves assumptions. Multivariate regression involves more assumptions; that means a second course on modeling. We can't ignore this, but we don't have time for it.

23

Teaching Confounding: Two Big Reasons Not To...

(1) Disrespect

(2) Prerequisites



**An open mind
is the prerequisite
to gaining
knowledge.**

Slide 24: These objections must be overcome before we can teach confounding in the intro course. But there are reasons **FOR** teaching confounding that offset these two objections.

#1: The Cornfield conditions can limit excessive skepticism. They set a minimum on the size of confounder that can negate or reverse an association. A butterfly flapping its wings in the Pacific can't reverse all the associations in the world – unless it is a very BIG butterfly. These are the conditions that Jerome Cornfield successfully used to reject Fisher's claim that the association between smoking and lung cancer might be confounded by genetics. See Schield (1999).

#2: When the predictor and confounder are binary, the regression assumptions are readily satisfied. The model is fully saturated. When the predictor and confounder are binary, a simple graphical technique is available so software is not needed. [I have taught this graphical technique for over 10 years in my Statistical Literacy course for students in non-quantitative majors.] My English, art and music majors can work problems involving confounding. See Schield (2006).

24

Teaching Confounding: Reasons To...

#1: The Cornfield conditions¹ set a minimum on the size confounder that can negate or reverse an association. Schield (1999). These conditions can offset excessive skepticism/cynicism.

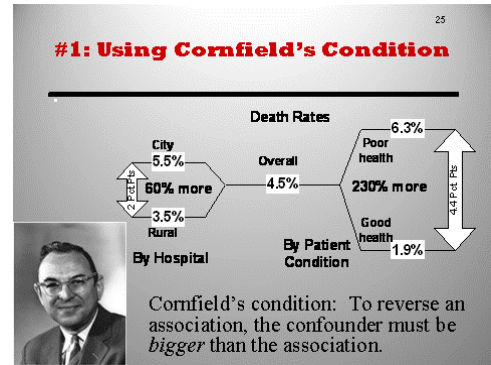
#2: When the predictor and confounder are binary, there are graphical techniques² that allow students to work problems without software and without a second course in regression. Schield (2006)

This material has been taught for over 10 years.

I'm going to show you examples of both of these along with references that you can pursue.

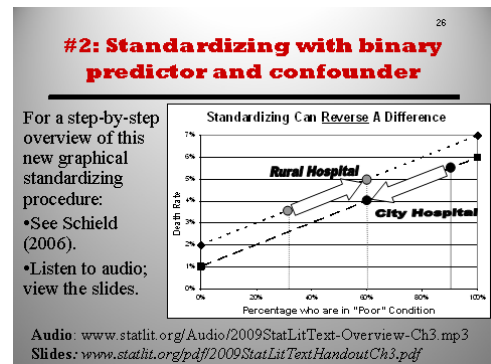
Slide 25: Here’s a way my students apply a Cornfield condition. Consider the patient death rate at two hospitals: City (5.5%) and Rural (3.5%). If you don’t want to die Rural hospital looks better. Rural probably has cookies and milk in the afternoon; City probably has more weird diseases.

But patient condition is certainly a confounder. The patient death rate is higher for those in poor health (6.3%) than for those in good health (1.9%). The patient health difference (4.4 percentage points) is bigger than the hospital difference (2 percentage points). Cornfield proved a necessary condition: that to negate or reverse an association, a confounder must be bigger than the association. See Schield (1999). My students can tell that patient condition could negate and even reverse this association between hospital and death rate.



Slide 26 shows a new graphical technique – standardizing – that controls for the influence of a binary confounder. Howard Wainer did statistical education a great service when he publicized this technique (Larry Lesser had identified it earlier). This technique takes a while to understand. See Schield (2006) and the references shown for more detail. The audio presentation takes you through step-by-step.

My students can see how an association of rates or percentages can be reversed by taking into account a confounder. This reversal is something they have never seen in any math course. And they can work problems.



Slide 27: My conclusion. Many – if not most – big-data users want causal explanations. In Business, this is what we call Business Intelligence. Modeling a time series, a change or a difference is a first step in seeing what caused those results.

Given this need, statistical educators must focus more on coincidence and confounding. Our students deserve a broader education than what we’ve been giving them.

We need to say more than “Association is not Causation”. [As Danny Kaplan says, we should let go of our “abstinence curriculum.”] We must introduce confounding along with the Cornfield conditions and standardization.

Most of the students taking introductory statistics major in the social sciences or professions: business sociology, social work, education and economics. Students in these disciplines deal primarily with observational data. Confounding is one of the biggest problems they have. Most of the textbooks in introductory statistics are silent on confounding (or mention it just briefly).

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.

[By using the absence of overlapping 95% confidence intervals as a sufficient condition for statistical significance, my students can see how statistical significance can be influenced by a confounder. This confounder influence on statistical significance is something every student should know. I don't know of any introductory statistics text that mentions this.]

I see the absence of confounding in our introductory textbooks [and the silence on how taking into account an extraneous factor can influence statistical significance] as professional negligence.

With big data, we need to deal with causation, confounding and coincidence. Now we have the tools to show how statistics is a foundational discipline in dealing with Big Data.

Thank you very much. I appreciate any comments or questions on this presentation. If you are interested in testing teaching materials based on these ideas, please let me know.

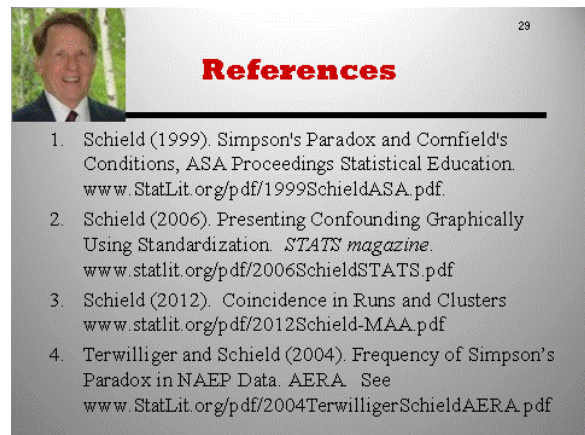
These slides are hosted at www.StatLit.org/pdf/2014-Schield-ECOTS-Slides.pdf

This paper is hosted at www.StatLit.org/pdf/2014-Schield-ECOTS.pdf

Slide 29: Here are my key references:

For more on statistical literacy, visit my website:

www.StatLit.org



Slide 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

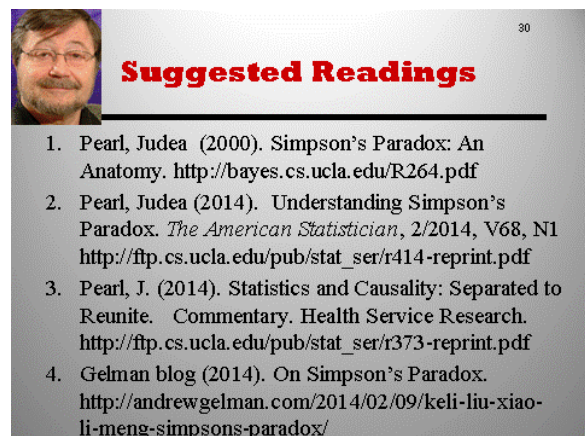
Slide 30: Here are some recommended readings.

Judea Pearl has two excellent articles on Simpson's Paradox. His 2014 paper is an absolute gem.

Statistics and Causality: Separated [at Birth] to Reunite [Today].

Hopefully, causality will be the hallmark of statistical education in the 21st century.

Gelman's blog is an interesting case study on how people talking on Simpson's Paradox can be talking at different levels: one at the modeling/prediction level; the other at the causal level.



Slide 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/>

ACKNOWLEDGEMENTS: I want to thank my colleagues, Marc Isaacson, Tom Burnham and Julie Naylor for their help and feedback on the many prior papers that went into this summary paper.

This paper is a combination of my slides, my verbal presentation and my subsequent additions [].