TWO BIG IDEAS IN TEACHING BIG DATA

Coincidence & Confounding

by
Milo Schield
Twin-Cities Chapter Meeting
March 19, 2014. Augsburg College
www.StatLit.org/pdf/2014-Schield-ASA-TCC-6up.pdf

Big Data and Big Ideas

Big data: "any data set in which *all* associations are statistically significant." [Schield definition]

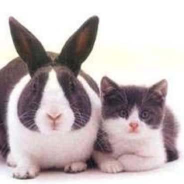
Leaving aside local experiments (A-B tests), it might seem that intro statistics – statistical significance – has little value with 'big data'.

In big data,

1. Coincidence is a bigger problem,
2. Confounding is often the #1 problem.

Coincidence?

.



The "Birthday" Problem: Chance of same birthday

Richard von Mises (1883-1953)

In a group of 28 people,
a birthday match (same month and day) is *expected*.

The "Birthday" Problem Math Answer

N!/[k!(N-k)!] combos of N things taken k at a time.

For k = 2, #combos = C = N(N-1)/2 ~ (N^2)/2

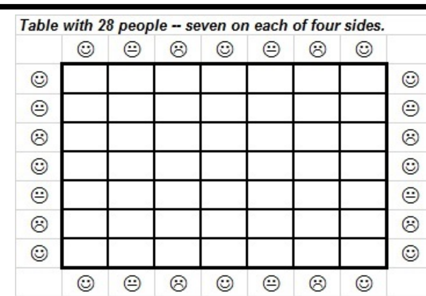
N ~ sqrt(2C). If C = 365, N ~ Sqrt(730) = 27.

Q. Are students convinced? No!!!

If the chance of an event is *p* and *p* =1/*n*,
then this event is "expected" in n trials.

Show students there are > 365 pairs w 28 people.

Consider a table



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

2014 Schield ASA TCC 7

Get Birthdays (Mn/Dy): Color cell with row-column match

Schield (2012)			RICHARD VON MISES' BIRTHDAY PROBLEM									V2b
			Press F9 for a new group of 28 people									
		Month	9	10	9	4	7	4	11			
		Day	24	3	26	26	18	28	6			
Month	Day		☺	😐	☹	☺	😐	☹	☺		Month	Day
4	9	☺								☺	2	15
8	10	😐								😐	7	18
2	20	☹								☹	5	19
6	30	☺								☺	8	15
2	22	😐								😐	5	9
6	17	☹								☹	7	25
1	15	☺								☺	4	11
			☺	😐	☹	☺	😐	☹	☺			
		Month	4	1	6	12	11	6	3			
		Day	7	27	26	4	11	18	9			

2014 Schield ASA TCC 8

Four Quadrants: 49 possible connections each

Schield (2011)			RICHARD VON MISES' BIRTHDAY PROBLEM							28 People	
		Month	10	11	11	9	4	7	6		
		Day	16	18	8	9	13	25	24		
Month	Day									Month	Day
8	20							1		7	25
10	29									8	16
4	11									11	6
3	3									11	29
1	3									8	3
3	30									3	24
10	28									1	15
		Month	5	2	6	2	1	7	5		
		Day	28	8	6	12	14	1	25		

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

2014 Schield ASA TCC 9

Top-to-Bottom & Left-to-Right: 49 connections each

Schield (2011)			RICHARD VON MISES' BIRTHDAY PROBLEM							28 People	
		Month	11	8	10	10	8	10	3		
		Day	19	3	28	17	27	29	5		
Month	Day					S				Month	Day
5	23									1	12
1	1									11	17
9	6									12	3
10	13									7	29
7	14									2	17
8	30									4	2
1	8									8	17
					N						
		Month	12	3	10	9	12	9	5		
		Day	24	6	17	19	1	20	29		

2014 Schield ASA TCC 10

Same-Edge (four): 21 connections each

Schield (2011)			RICHARD VON MISES' BIRTHDAY PROBLEM								28 People	
		Month	3	2	2	3	9	3	5			
		Day	4	5	9	29	20	5	20			
Month	Day										Month	Day
6	22									E	4	1
10	8										7	10
5	5										3	26
11	23										3	10
3	27									E	4	1
10	2										9	8
2	21										5	7
		Month	8	1	10	12	9	5	5			
		Day	18	6	11	9	3	26	19			

2014 Schield ASA TCC 11

Connections and Chance

Pairs	GROUP	Details
196	Quadrants 1-4	49 pairs each
49	Left-to-Right	
49	Top-to-Bottom	
84	Within each side	21 pairs each
378	TOTAL	

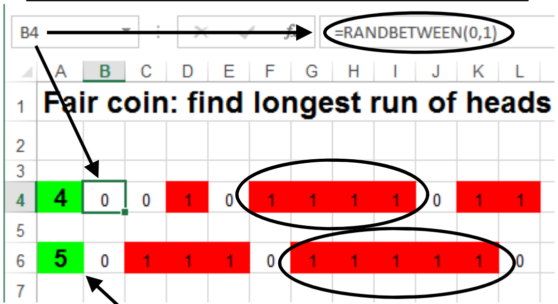
A *preselected* birthday match has one chance in 365.
In a group of 28, we have 378 pairs: (N-1)(N/2).
A *somewhere* match is expected – > 50% of the time.

Coincidence: Flipping a fair coin Getting a "run" of heads

Conjecture:
The longer
the run,
the more unlikely
the outcome.

Empirical test

Flip coins in rows. 1=Heads Red = Run of heads.



Green: Length of longest run in that row

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

15

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

Question is ambiguous! Doesn't state context!

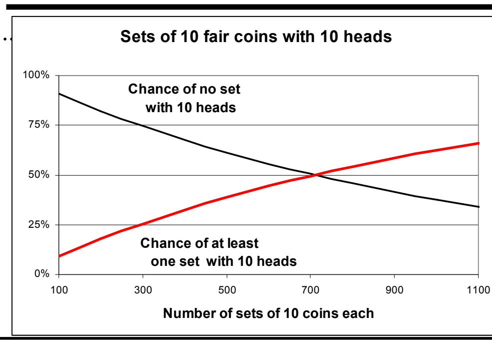
1. Chance of 10 heads on **the next 10 flips**?
   p = 1/2; k = 10.
   P = p^k = (1/2)^10 = one chance in 1,024

2. Chance of at least one run of 10 heads **somewhere** when flipping 1,024 sets* of 10 coins each? At least 50%

* or (conjecture) when flipping 1,033 coins: 1/p + k-1.

16

Coincidence increases as data size increases



17

Law of Coincidence

Law of Very-Large Numbers (Qualitative):
The unlikely is almost certain given enough tries

Law of Expected Values:

Consider N tries with events having one chance in N.

* One event 'expected' in N tries

* Chance of at least one > 50%

18

Second Big Idea: Confounding

Big data will force statistical education to deal with causation in observational studies.

1. Most big data are observational.
2. Most big data users want to use associations as evidence for causation.
3. Confounding is the #1 problem.
4. 'Confound', 'predict' and explain', will need clarification.

19

Confounding: Two definitions

Confounder (math): **Associated/observational**

Any factor associated with the predictor (independent) and with the outcome (dependent) in an association.

Confounder (Epidemiological): **Causal/experimental**

Any factor associated with the predictor (independent) and with the outcome (dependent) in an association:

- that is *not caused* by the predictor, and
- that has *a causal influence* on the outcome.

20

Prediction: Two definitions

Prediction (math): **Associated/observational**

Modelled result assuming none of the factor levels are set by a researcher.

Prediction (Business): **Causal/Experimental**

Modelled results based on factor levels that could be set by a researcher.

21

Explain: Two definitions

Explain (math): **Associated/observational**

How much of the outcome variation is *associated with* or *attributable to* a given factor.*

Explains (Business): **Causal/Experimental**

How much of the outcome variation is *a result of* or *caused by* a factor.*

* 'Due to' and 'because of' are "in-between"

22

Common Confusions

Among adult men:

1. Weight and height are positively correlated.
2. Those who are heavier are generally taller than those who are thinner.

3. As weight increases, height increases.
4. For every extra 5#, height increases by 1 inch
5. If you gain weight, you will grow taller.

23

Ambiguity in "Explains"

For every 5# increase in weight in adult men, height increases by 1 inch.

Does the five pound increase in weight "explain" the one inch increase in height?

- Yes, if explain means "is associated with": we shift focus from light-weight men to heavy-weight men at a given time.
- No, if explains means "causes": we increase the weight of individual men over time.

24

Multivariate Analysis Predict vs. "Explain"

Step	1	2	3	
Constant	$80,000	$78,000	$58,000	
Baths	$39,000	$36,000	$15,000	per bath
Acres		$7,500	$7,500	per acre
Area			$33	per sq. foot
R-sq	44%	60%	68%	

Predict/observe: accuracy ↑ as factors ↑

#3: Each extra bath *explains* a $15K ↑ in value.

Predict/causal: If a bathroom is added, the house value is expected to ↑ by $15K.

25

Modeling: What to Take into Account

Consider modeling the outcome in this causal diagram:

Predictor → Confounder → Outcome

Kaplan: Model Outcome on Predictor
Schield: Model Outcome on Predictor and Confounder

1. Who is right?
2. Can both be right? YES!!!

Schield in predicting; Kaplan in causal explaining.

26

Causation & Simpson's Paradox

Simpson's paradox is not a paradox in prediction.

Simpson's paradox is only a paradox in forming a causal explanation or conclusion.

In a prediction the signs and sizes of the coefficients are all but irrelevant. R-sq is what counts.

In a causal explanation, the size and sign of the coefficients matter. R-sq is all but irrelevant.

27

Conclusion

Many – if not most – big-data users want causal explanations and causal predictions.

Math-stats can help us explain why coincidence increases as the size of the data increases.

Mathematics doesn't study causation. There is no mathematical operator or operation for causes.

Statistics education must say more about causation than simply saying "Association is not Causation."

28

Recommendations

1. Schield (2011) Coincidence in Runs and Clusters www.statlit.org/pdf/2012Schield-MAA.pdf
2. Pearl (2000). Simpson's Paradox: An Anatomy. http://bayes.cs.ucla.edu/R264.pdf
3. Pearl (2009), Causal inference in statistics. http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf
4. Gelman blog (2014). On Simpson's Paradox. http://andrewgelman.com/2014/02/09/keli-liu-xiao-li-meng-simpsons-paradox/

TWO BIG IDEAS IN TEACHING BIG DATA

Coincidence & Confounding

by
Milo Schield

Twin-Cities Chapter Meeting
March 19, 2014. Augsburg College

www.StatLit.org/pdf/2014-Schield-ASA-TCC-6up.pdf

Big Data and Big Ideas

Big data: "any data set in which *all* associations are statistically significant." [Schield definition]

Leaving aside local experiments (A-B tests), it might seem that intro statistics – statistical significance – has little value with 'big data'.

In big data,

1. Coincidence is a bigger problem,
2. Confounding is often the #1 problem.

Coincidence?

The "Birthday" Problem: Chance of same birthday

Richard von Mises (1883-1953)

In a group of 28 people,
a birthday match (same month and day) is *expected*.

The "Birthday" Problem
Math Answer

N!/[k!(N-k)!] combos of N things taken k at a time.

For k = 2, #combos = C = N(N-1)/2 ~ (N^2)/2

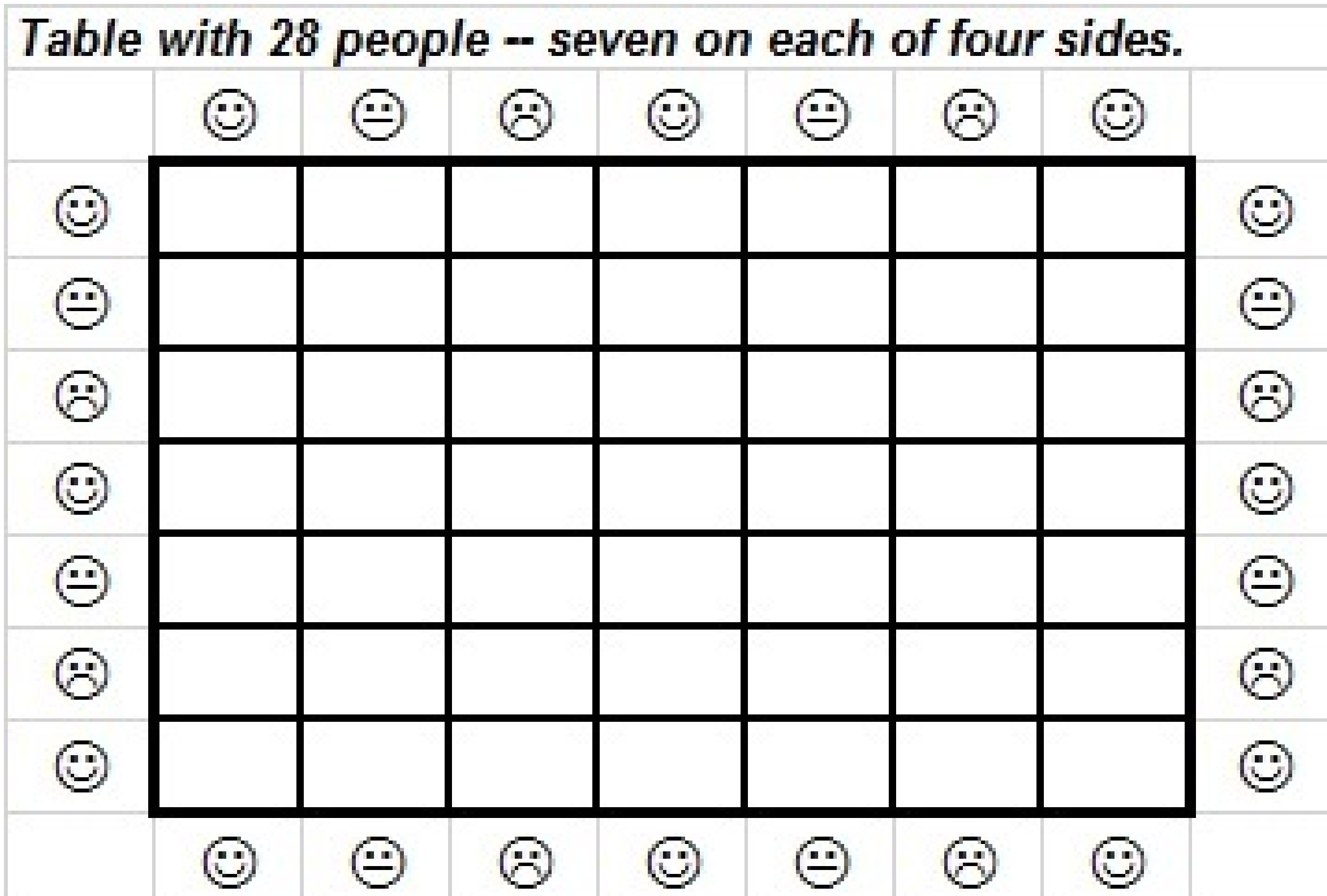
N ~ sqrt(2C). If C = 365, N ~ Sqrt(730) = 27.

Q. Are students convinced? No!!!

If the chance of an event is *p* and *p* =1/*n*,
then this event is "expected" in n trials.

Show students there are > 365 pairs w 28 people.

Consider a table



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

Get Birthdays (Mn/Dy): Color cell with row-column match

Schield (2012)			RICHARD VON MISES' BIRTHDAY PROBLEM									V2b
			Press F9 for a new group of 28 people									
		Month	9	10	9	4	7	4	11			
		Day	24	3	26	26	18	28	6			
Month	Day		☺	😐	☹	☺	😐	☹	☺		Month	Day
4	9	☺								☺	2	15
8	10	😐								😐	7	18
2	20	☹								☹	5	19
6	30	☺								☺	8	15
2	22	😐								😐	5	9
6	17	☹								☹	7	25
1	15	☺								☺	4	11
			☺	😐	☹	☺	😐	☹	☺			
		Month	4	1	6	12	11	6	3			
		Day	7	27	26	4	11	18	9			

Four Quadrants: 49 possible connections each

Schield (2011)			RICHARD VON MISES' BIRTHDAY PROBLEM							28 People	
		Month	10	11	11	9	4	7	6		
		Day	16	18	8	9	13	25	24		
Month	Day									Month	Day
8	20							1		7	25
10	29									8	16
4	11									11	6
3	3									11	29
1	3									8	3
3	30									3	24
10	28									1	15
		Month	5	2	6	2	1	7	5		
		Day	28	8	6	12	14	1	25		

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

Top-to-Bottom & Left-to-Right: 49 connections each

Schield (2011)			RICHARD VON MISES' BIRTHDAY PROBLEM								28 People
		Month	11	8	10	10	8	10	3		
		Day	19	3	28	17	27	29	5		
Month	Day					S				Month	Day
5	23									1	12
1	1									11	17
9	6									12	3
10	13									7	29
7	14									2	17
8	30									4	2
1	8									8	17
					N						
		Month	12	3	10	9	12	9	5		
		Day	24	6	17	19	1	20	29		

Same-Edge (four): 21 connections each

Schield (2011)			RICHARD VON MISES' BIRTHDAY PROBLEM							28 People	
		Month	3	2	2	3	9	3	5			
		Day	4	5	9	29	20	5	20			
Month	Day										Month	Day
6	22									E	4	1
10	8										7	10
5	5										3	26
11	23										3	10
3	27									E	4	1
10	2										9	8
2	21										5	7
		Month	8	1	10	12	9	5	5			
		Day	18	6	11	9	3	26	19			

Connections and Chance

Pairs	GROUP	Details
196	Quadrants 1-4	49 pairs each
49	Left-to-Right	
49	Top-to-Bottom	
84	Within each side	21 pairs each
378	TOTAL	

A *preselected* birthday match has one chance in 365.

In a group of 28, we have 378 pairs: (N-1)(N/2).

A *somewhere* match is expected – > 50% of the time.

Coincidence: Flipping a fair coin
Getting a “run” of heads

Conjecture:

The longer

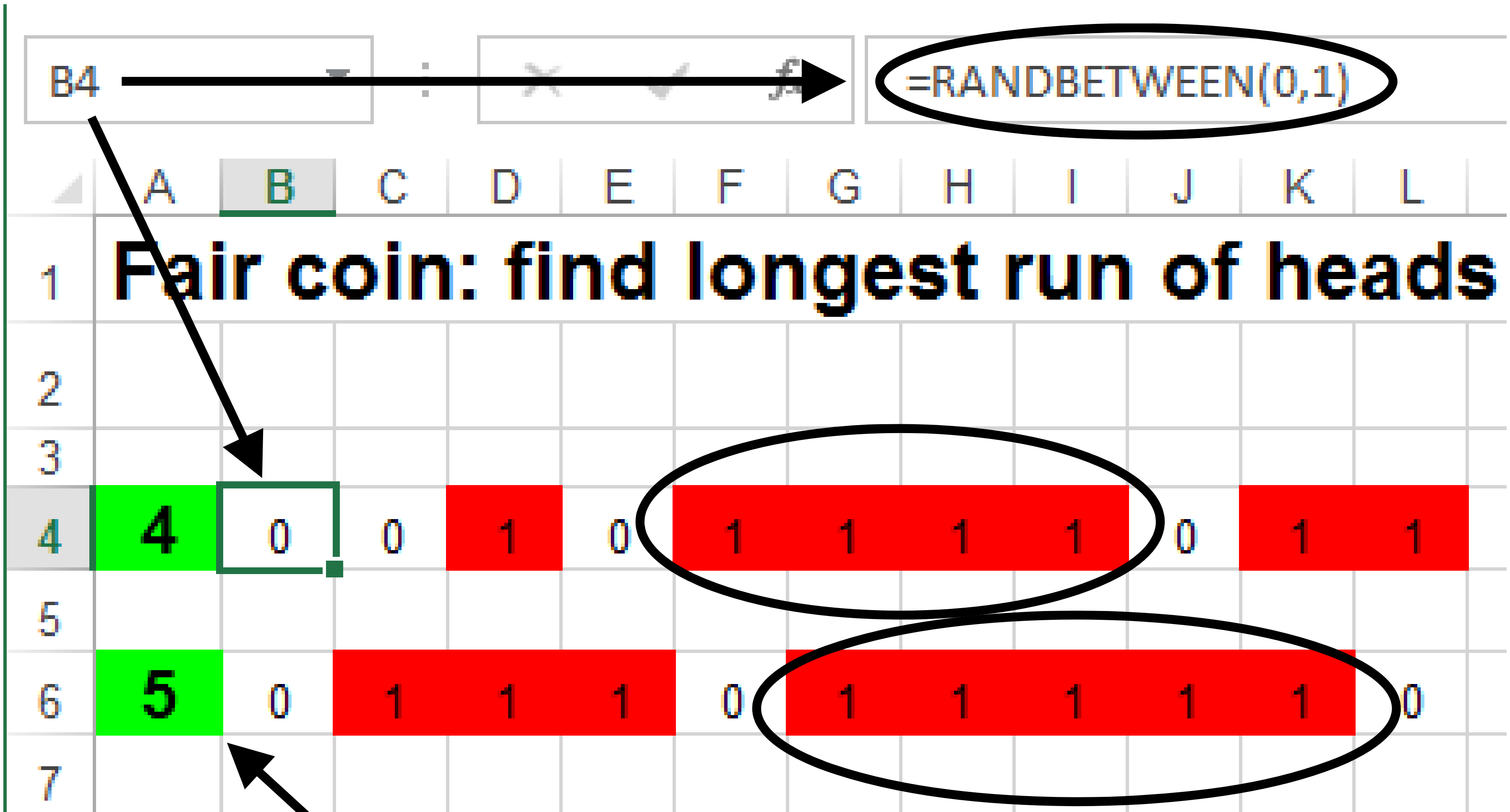
the run,

the more unlikely

the outcome.

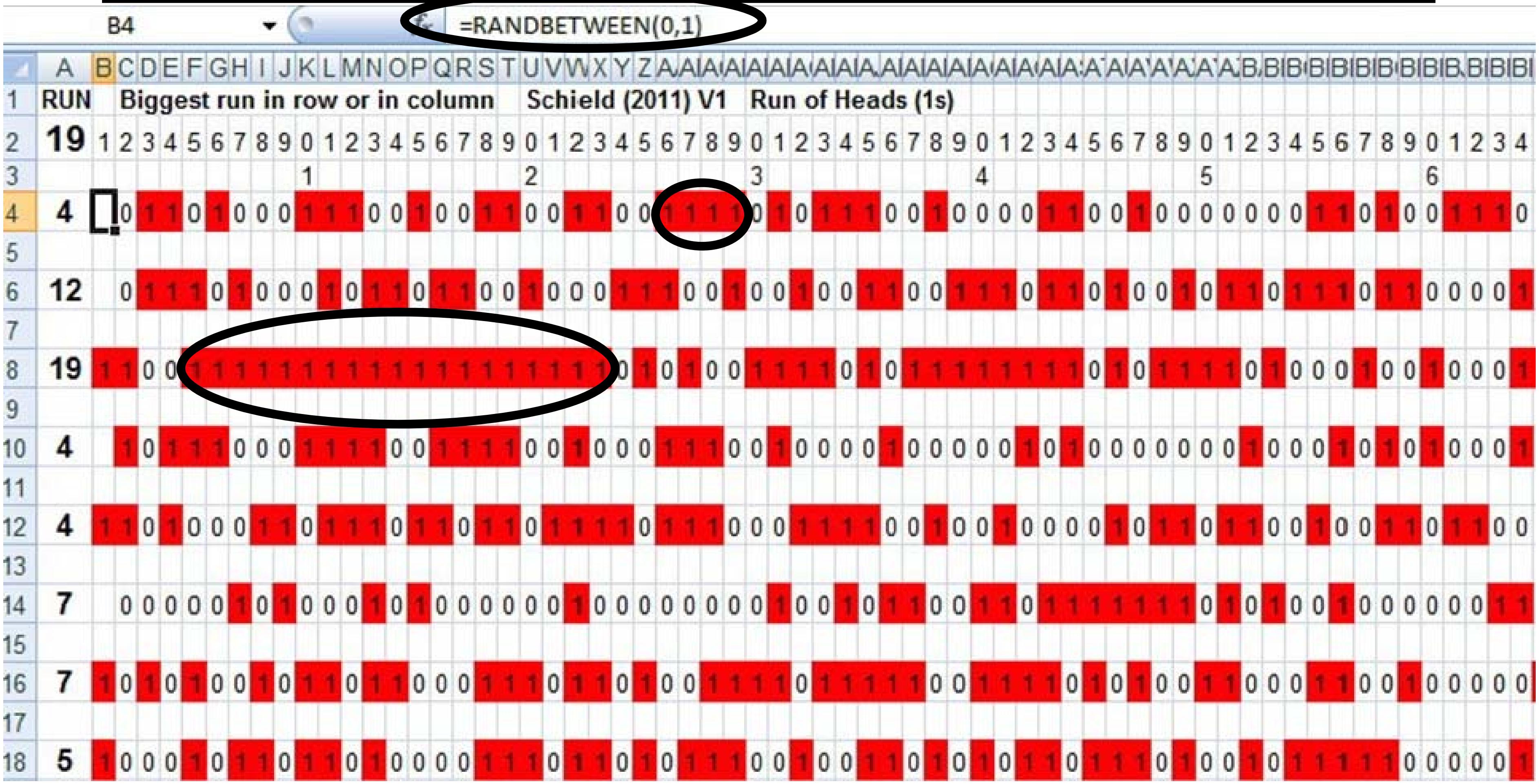
Empirical test

Flip coins in rows. 1=Heads
Red = Run of heads.



Green: Length of longest run in that row

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

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

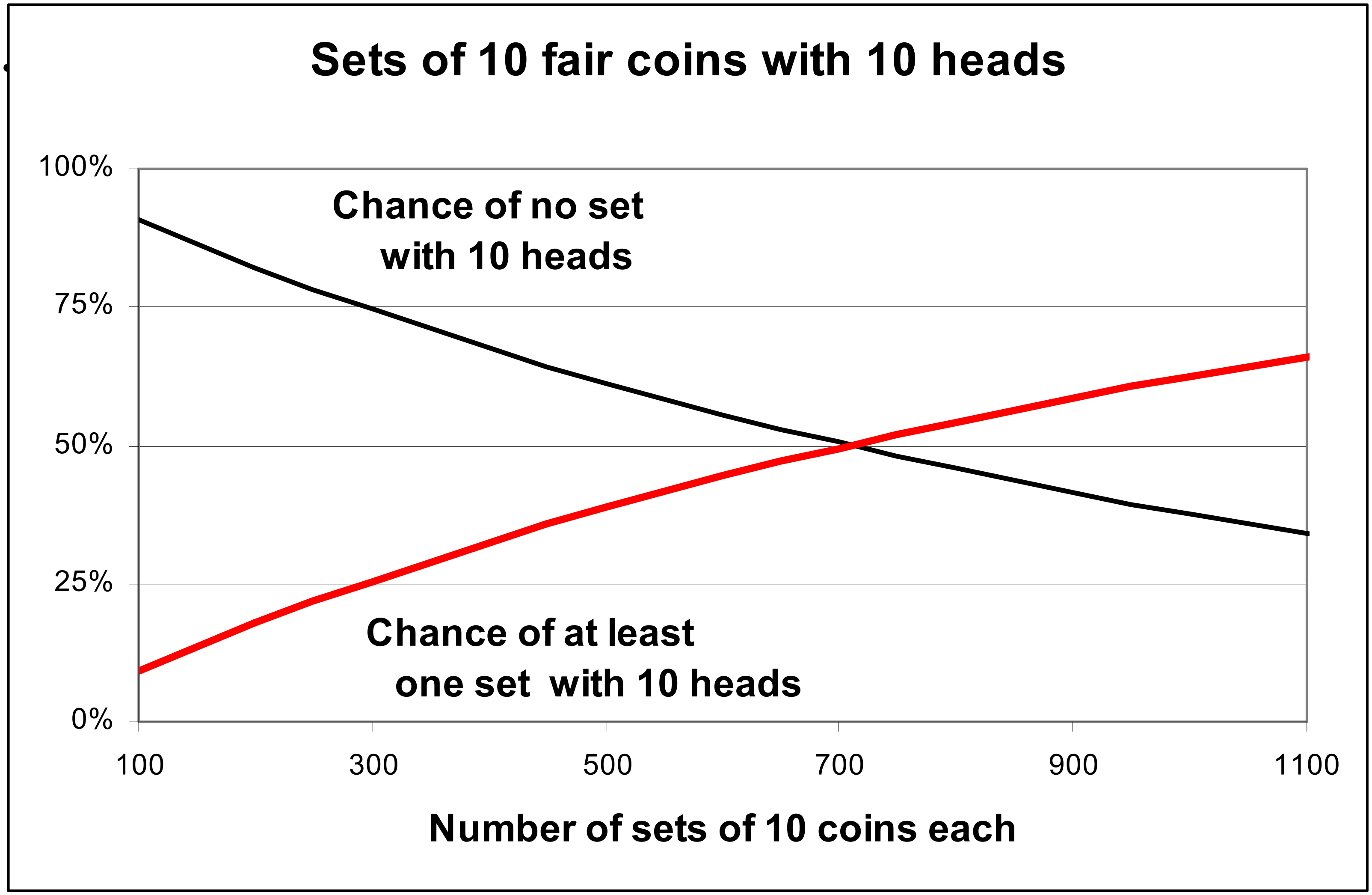
Question is ambiguous! Doesn't state context!

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. Chance of at least one run of 10 heads **somewhere** when flipping 1,024 sets* of 10 coins each? At least 50%

* or (conjecture) when flipping 1,033 coins: 1/p + k-1.

Coincidence increases as data size increases



Law of Coincidence

Law of Very-Large Numbers (Qualitative):
The unlikely is almost certain given enough tries

Law of Expected Values:

Consider N tries with events having one chance in N.

* One event 'expected' in N tries
* Chance of at least one > 50%

Second Big Idea: Confounding

Big data will force statistical education to deal with causation in observational studies.

1. Most big data are observational.
2. Most big data users want to use associations as evidence for causation.
3. Confounding is the #1 problem.
4. ‘Confound’, ‘predict’ and explain’, will need clarification.

Confounding: Two definitions

Confounder (math): **Associated/observational**

Any factor associated with the predictor (independent) and with the outcome (dependent) in an association.

Confounder (Epidemiological): **Causal/experimental**

Any factor associated with the predictor (independent) and with the outcome (dependent) in an association:

- that is *not caused* by the predictor, and
- that has *a causal influence* on the outcome.

Prediction: Two definitions

Prediction (math): **Associated/observational**

Modelled result assuming none of the factor levels are set by a researcher.

Prediction (Business): **Causal/Experimental**

Modelled results based on factor levels that could be set by a researcher.

Explain:
Two definitions

Explain (math): **Associated/observational**

How much of the outcome variation is *associated with* or *attributable to* a given factor.*

Explains (Business): **Causal/Experimental**

How much of the outcome variation is *a result of* or *caused by* a factor.*

* 'Due to' and 'because of' are "in-between"

Common Confusions

Among adult men:

1. Weight and height are positively correlated.
2. Those who are heavier are generally taller than those who are thinner.

3. As weight increases, height increases.
4. For every extra 5#, height increases by 1 inch
5. If you gain weight, you will grow taller.

Ambiguity in "Explains"

For every 5# increase in weight in adult men, height increases by 1 inch.

Does the five pound increase in weight "explain" the one inch increase in height?

- Yes, if explain means "is associated with": we shift focus from light-weight men to heavy-weight men at a given time.
- No, if explains means "causes": we increase the weight of individual men over time.

Multivariate Analysis
Predict vs. "Explain"

Step	1	2	3	
Constant	\$80,000	\$78,000	\$58,000	
Baths	\$39,000	\$36,000	\$15,000	per bath
Acres		\$7,500	\$7,500	per acre
Area			\$33	per sq. foot
R-sq	44%	60%	68%	

Predict/observe: accuracy ↑ as factors ↑

#3: Each extra bath *explains* a \$15K ↑ in value.

Predict/causal: If a bathroom is added, the house value is expected to ↑ by \$15K.

Modeling:
What to Take into Account

Consider modeling the outcome in this causal diagram:

Predictor → Confounder → Outcome

Kaplan: Model Outcome on Predictor
Schield: Model Outcome on Predictor and Confounder

1. Who is right?

2. Can both be right? YES!!!

   Schield in predicting; Kaplan in causal explaining.

Causation & Simpson's Paradox

Simpson's paradox is not a paradox in prediction.

Simpson's paradox is only a paradox in forming a causal explanation or conclusion.

In a prediction the signs and sizes of the coefficients are all but irrelevant. R-sq is what counts.

In a causal explanation, the size and sign of the coefficients matter. R-sq is all but irrelevant.

Conclusion

Many – if not most – big-data users want causal explanations and causal predictions.

Math-stats can help us explain why coincidence increases as the size of the data increases.

Mathematics doesn't study causation. There is no mathematical operator or operation for causes.

Statistics education must say more about causation than simply saying "Association is not Causation."

Recommendations

1. Schield (2011) Coincidence in Runs and Clusters www.statlit.org/pdf/2012Schield-MAA.pdf
2. Pearl (2000). Simpson's Paradox: An Anatomy. http://bayes.cs.ucla.edu/R264.pdf
3. Pearl (2009), Causal inference in statistics. http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf
4. Gelman blog (2014). On Simpson's Paradox. http://andrewgelman.com/2014/02/09/keli-liu-xiao-li-meng-simpsons-paradox/