V0 2016 IASE 1

B: Teaching Confounding and Multivariate Thinking

Milo Schield, Augsburg College
Member: International Statistical Institute
US Rep: International Statistical Literacy Project
VP. National Numeracy Network

IASE Roundtable in Berlin
July 20, 2016
www.StatLit.org/pdf/2016-Schield-IASE-2Slides.pdf

V0 2016 IASE-2 2

GAISE 2016: Two New Emphases

a. **Teach statistics as an investigative process of problem-solving and *decision making*.**
   - Statistics is a problem-solving and decision-making process, not a collection of formulas and methods.

b. **Give students experience in *multivariable thinking***
   - The world is a tangle of complex problems with inter-related factors. Lets show students how to explore relationships among many variables

V0 2016 IASE-2 3

GAISE 2016 Add Multivariable Thinking

- give "students experience with multivariable thinking"
- understand "the possible impact of ... *confounding*"
- See how "a third variable can change our understanding"
- Help students "identify *observational studies*"
- teach multivariate thinking "in stages" and
- use "simple approaches (such as stratification)"

This change is HUGE! It may be the biggest content change since dropping combinations in the 1980s.

V0 2016 IASE-2 4

GAISE 2016 Appendix B: Observational Data

Multivariable thinking is critical to make sense of the *observational data* around us. The real world is complex and can't be described well by one or two variables. [Italics added]

V0 2016 IASE-2 5

GAISE 2016 Confounding

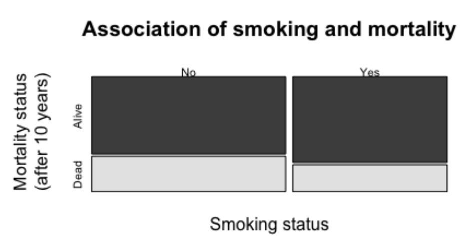
"The 2014 ASA guidelines for undergraduate programs in statistics recommend that students obtain a clear understanding of principles of statistical design and tools to assess and account for *the possible impact of other measured and unmeasured confounding variables* (ASA, 2014)."

http://www.amstat.org/education/gaise/collegeupdate/GAISE2016_DRAFT.pdf

V0 2016 IASE-2 6

Show Multivariable #1: Ekisogram

Show probabilities as areas:

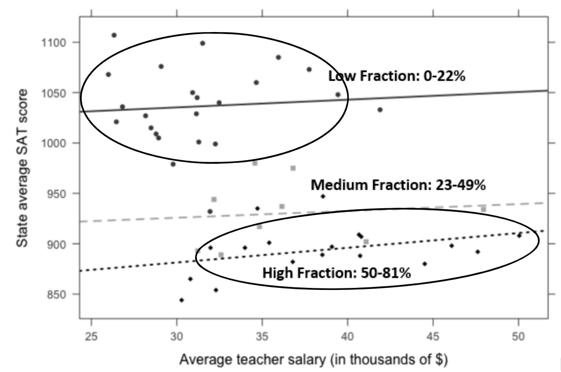


This mosaic plot doesn't work well for me.

V0 2016 IASE-2 7

Show Multivariable: #2: XY Plot (2 factors)

State Average SAT Score by Average Teacher Salary
Series: Fraction of Students that took the SAT



V0 2016 IASE-2 9

#2 Show Multivariable: Confounder is Too Complex

This method models separate series in that same XY plot. The confounder: percentage of students in the state that took the SAT.

- Consider the "low-fraction" states in the upper-left corner. Most students in the Middle states take the ACT – not the SAT. Only the best "middle" students take the SAT in applying to colleges on the East or West coast. In the "middle" teacher salaries are lower.
- Consider the "high fraction" states in the lower-right corner. Most students on the East and West coast take the SAT. These students include all students: best, middle and below-average so their average SAT is lower. On the coasts, teacher salaries are higher.

Controlling for the percentage taking the SAT changes the association between teacher salaries and average student scores.

V0 2016 IASE-2 10

#3 Show Multivariable Regression X-Y Output

Scottish Hill Races (Time in seconds)

Response variable is: Women's Record
R squared = 85.2% R squared (adjusted) = 84.9%
s = 1126 with 70-2 = 68 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	320.528	222.2	1.44	0.1537
Climb	1.755	0.088	19.8	< 0.0001

Assume: All modelling assumptions are satisfied
Assume: All slope coefficients are statistically significant.
http://www.scottishhillracing.co.uk/

V0 2016 IASE-2 11

#3 Show Multivariate: Regression X1-X2-Y Output

Scottish Hill Races (Time in seconds)

Response variable is: Women's Record
R squared = 97.5% R squared (adjusted) = 97.4%
s = 468.0 with 70 - 3 = 67 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	-497.656	102.8	-4.84	< 0.0001
Distance	387.628	21.45	18.1	< 0.0001
Climb	0.852	0.0621	13.7	< 0.0001

Controlling for Distance decreases Climb coefficient from 1.755 to 0.852; increases R^2 from 85% to 97%.

V0 2016 IASE-2 12

2016 GAISE Appendix B: Closing Thoughts (1)

"Multivariable thinking is critical to make sense of the observational data around us. This type of thinking might be introduced in stages":

1. Learn to identify observational studies
2. Why randomized assignment … improves things
3. Wary: cause-effect conclusions from observational data
4. Consider – *and explain* -- confounding factors
5. Simple approaches (stratification) to *show* confounding

http://www.amstat.org/education/gaise/collegeupdate/GAISE2016_DRAFT.pdf

V0 2016 IASE-2 13

2016 GAISE Appendix B Closing Thoughts (2)

"If students do not have exposure to simple tools for disentangling complex relationships, they may dismiss statistics as an old-school discipline only suitable for small sample inference of randomized studies."

"This report recommends that students be introduced to multivariable thinking, preferably early in the introductory course and not as an afterthought at the end of the course."

V0 2016 IASE-2 14

GAISE 2016 Deletions

Making Room: What might be voted off the island?
(or at least slimmed down)

- Probability theory.
- Constructing plots by hand.
- Basic statistics by hand.
- Drills with *z*-, *t*-, χ2, and F-tables.
- Advanced training on a statistics software program.

V0 2016 IASE-2 15

Five Other Methods for Presenting Confounding

A. Show confounding
1. Stratification using 2x2 averages tables
2. Stratification using 2x2 rate tables

B. Explain confounding
1. Mixture Displays
2. Wainer diagrams
3. Reverse-engineering rate tables

V0 2016 IASE-2 16

A1: Show Confounding: Stratified 2x2 Averages Table

At age 20, the average male-female weight difference is:
27 pounds [156 – 129] Average cells have grey fill.

Ave Weight	Ht=64"	Ht=70"
FEMALE	129	142
MALE		156

14 pounds [156-142] **after controlling for height**.

* www.cdc.gov/growthcharts/html_charts/bmiagerev.htm

V0 2016 IASE-2 17

A2: Show Confounding: Stratified 2x2 Rate Tables

Death Rates by Group

DIED	YOUNG	OLD	TOTAL
NON SMOKER	12%	86%	31%
SMOKER	18%	88%	24%
TOTAL	15%	86%	28%

Non-smokers are more likely to die than smokers

Within Young (and within Old), the reverse is true.

V0 2016 IASE-2 18

Problem with "Showing" Confounding

1. Do these visualizations "explain" confounding?
2. Can students use these devices to work problems with numerical answers?
3. Will any of this be on the final?

If all three answers are "No", teachers are unlikely to spend much time showing multivariable thinking.

Maybe during the last class before the final ☺

V0 2016 IASE-2 19

B1. Explain Confounding: Explicit Mixture Displays

	Year 1		Year 2		
	Number	Score	#	Score	
Disavantaged	10	80%	50	81%	⇧
Advantaged	90	90%	50	91%	⇧
TOTAL	100	89%	100	86%	⇩

After Year 1, other disadvantaged student switch to this teacher increasing their prevalence from 10% to 50%.

Teacher's scores: Better for each group; worse overall.

Explanation: "It's the mix"

V0 2016 IASE-2 20

B2. Explain Confounding: Wainer's Standardization

Wainer (2004) introduced a graphical technique that controlled for the influence of a binary confounder.

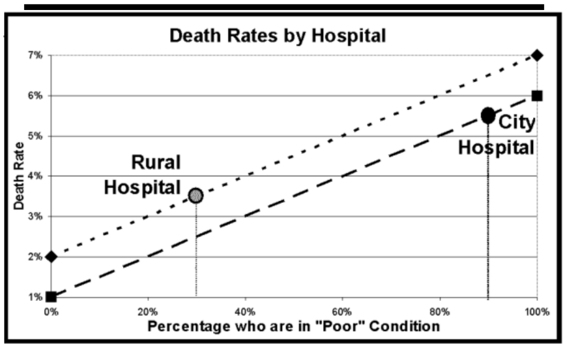
It requires minimal math and is visually intuitive.

My music and art majors find this graph easy to read. They can work problems with numerical answers.

For the origin (1986) and details, see
> Tan (2012): www.statlit.org/pdf/2012-Tan-Simpsons-Paradox.pdf
> Schield (2006): www.statlit.org/pdf/2006SchieldSTATS.pdf.

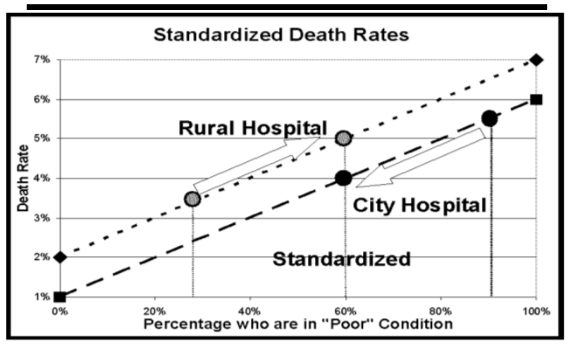
V0 2016 IASE-2 21

#B2: Wainer Diagrams Simpson's Paradox: It's the Mix



V0 2016 IASE-2 22

Simpson's Paradox: It's the Mix Standardize: Common Mixture



V0 2016 IASE-2 23

B3. Explain Confounding: Reverse-Engineer Rate Tables

DIED	YOUNG	OLD	TOTAL
NON SMOKER	12%	86%	31%
SMOKER	18%	88%	24%
TOTAL	15%	86%	28%

74% of top row are young; 90% of Row 2 are young.
82% of Row 3 are young; standardize top 2 with 82% young
Non-smoker standard death rate: 25% (0.82*12+0.18*86)
Smoker standardized death rate: 31% (0.82*18+0.18*88)
Standardized death rate for smokers > than for non-smokers

V0 2016 IASE 24

Why Statistical Educators Won't Teach Confounding

1. Students will have less trust in statistics if any confounder can reverse any association
2. Statisticians are not subject-matter experts
3. Emphasizes inductive/hypothetical thinking
4. Co-variation and sufficiency are math; confounding and causation are not.
5. "Association is not causation". K. Pearson: Causation is "a fetish amidst the inscrutable arcana of modern science"

V0 2016 IASE 25

"Less Trust" vs. Cornfield Conditions

1950s: Fisher said that the smoking-death (10X) association might be confounded by genetics (3X).

Cornfield proved that to nullify (or reverse) this association, the confounder must exceed 10X.

"Cornfield's minimum effect size is as important to observational studies as is the use of randomized assignment to experimental studies." Schield (1999)

Schield (1999) www.statlit.org/pdf/1999SchieldASA.pdf

V0 2016 IASE-2 26

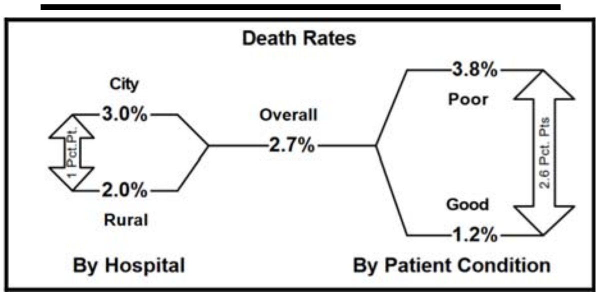
Stratification Two-Way Half Tables

Patient Died	"Good"	"Poor"	TOTAL
City Hospital	1%	6%	5.5%
Rural Hospital	2%	7%	3.5%
TOTAL	1.5%	6.5%	

City patient is 2 pts more likely to die that a Rural patient.
Poor patient is 5 pts more likely to die than a Good patient.
Association with Outcome: Confounder > Predictor

V0 2016 IASE 27

Cornfield Condition for Nullification or Reversal



Schield (1999) based on realistic data

V0 2016 IASE 28

Cornfield Condition for Nullification or Reversal

An association is nullified or reversed ***only if***

- *confounder (patient condition) has a stronger association with the outcome (death) than does the predictor (hospital).*
- *predictor (hospital) has a stronger association with the confounder (patient condition) than with the outcome (death).*

V0 2016 IASE-2 29

Teaching Confounding

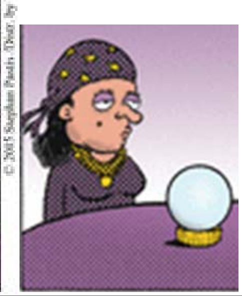
The bigger the effect size, the less likely a confounder can negate or reverse and observed association.

Effect Sizes:

- 10X: Smoking and death from lung cancer
- 1.3X: Second hand smoke and death

30

Confounded



V0 2016 IASE 1

B: Teaching Confounding and Multivariate Thinking

Milo Schield, Augsburg College

Member: International Statistical Institute

US Rep: International Statistical Literacy Project

VP. National Numeracy Network

IASE Roundtable in Berlin

July 20, 2016

www.StatLit.org/pdf/2016-Schield-IASE-2Slides.pdf

GAISE 2016: Two New Emphases

a. **Teach statistics as an investigative process of problem-solving and *decision making*.**
   - Statistics is a problem-solving and decision-making process, not a collection of formulas and methods.

b. **Give students experience in *multivariable thinking***
   - The world is a tangle of complex problems with inter-related factors. Lets show students how to explore relationships among many variables

V0 2016 IASE-2 3

GAISE 2016
Add Multivariable Thinking

- give "students experience with multivariable thinking"
- understand “the possible impact of ... *confounding*"
- See how "a third variable can change our understanding"
- Help students "identify *observational studies*"
- teach multivariate thinking "in stages" and
- use "simple approaches (such as stratification)”

This change is HUGE! It may be the biggest content change since dropping combinations in the 1980s.

GAISE 2016 Appendix B: Observational Data

Multivariable thinking is critical to make sense of the *observational data* around us. The real world is complex and can't be described well by one or two variables. [Italics added]

GAISE 2016
Confounding

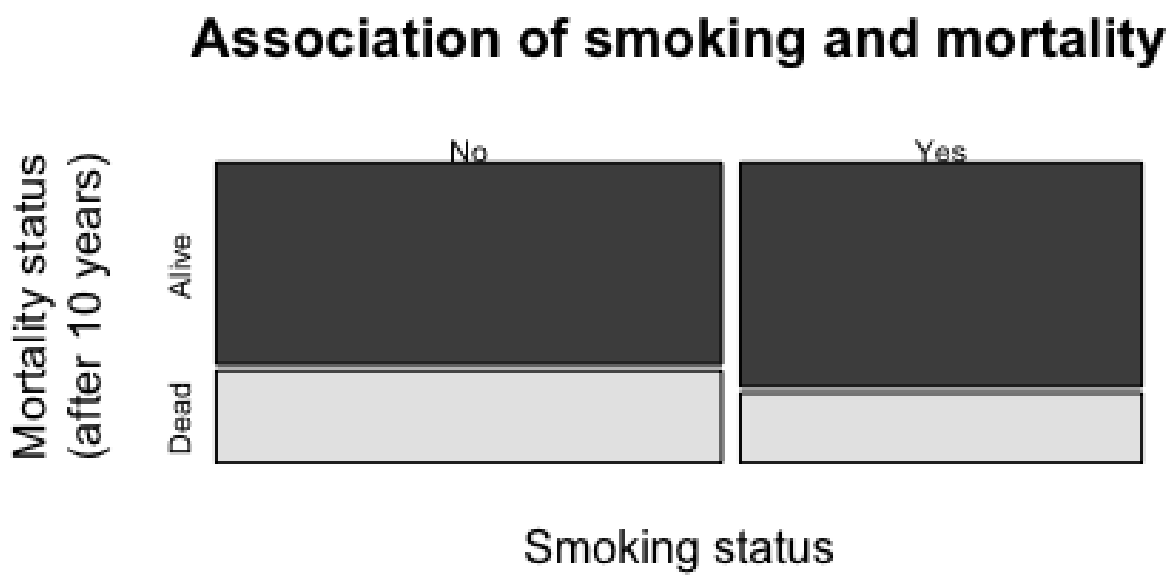
“The 2014 ASA guidelines for undergraduate programs in statistics recommend that students obtain a clear understanding of principles of statistical design and tools to assess and account for *the possible impact of other measured and unmeasured confounding variables* (ASA, 2014).“

http://www.amstat.org/education/gaise/collegeupdate/GAISE2016_DRAFT.pdf

V0 2016 IASE-2 6

Show Multivariable #1: Ekisogram

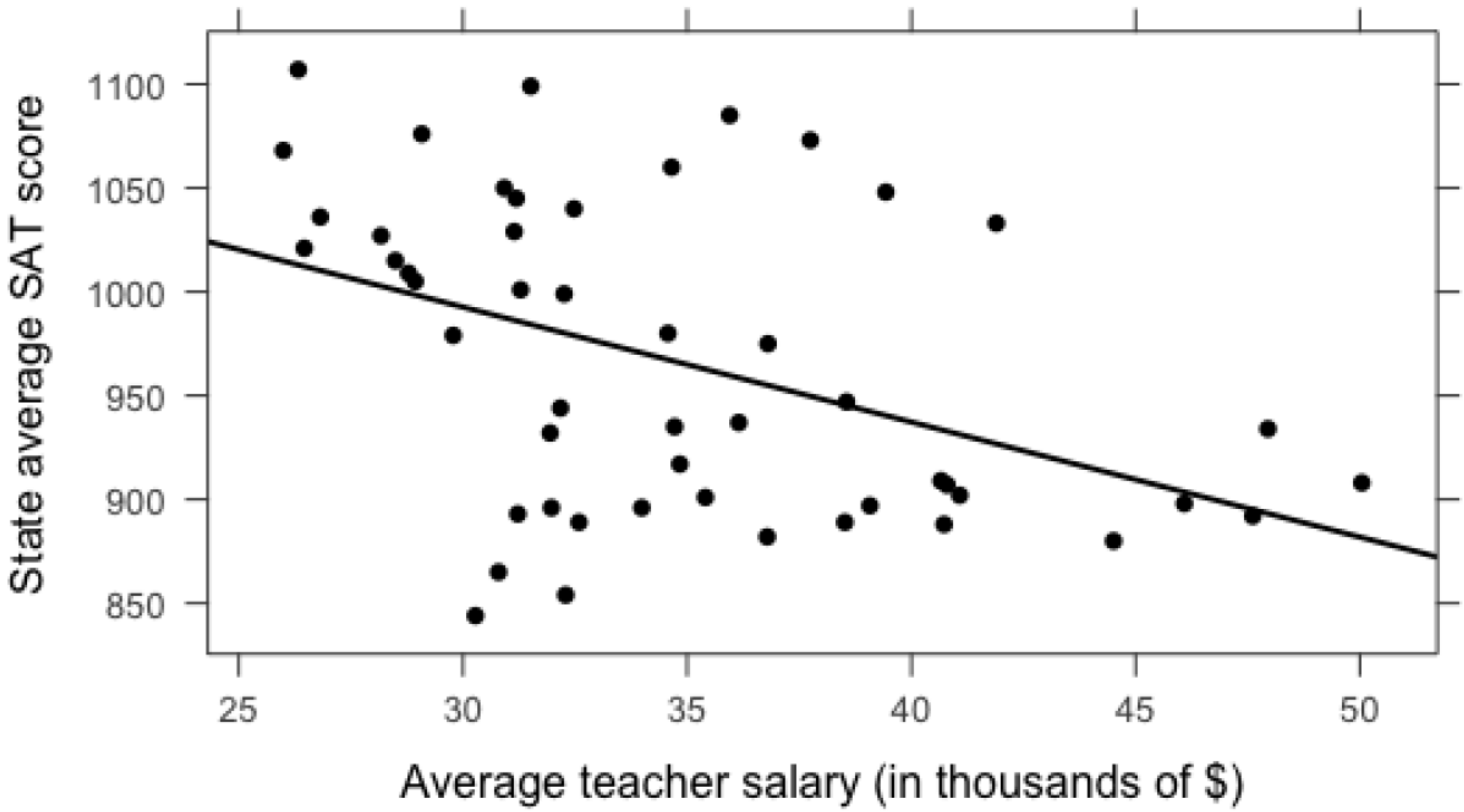
Show probabilities as areas:

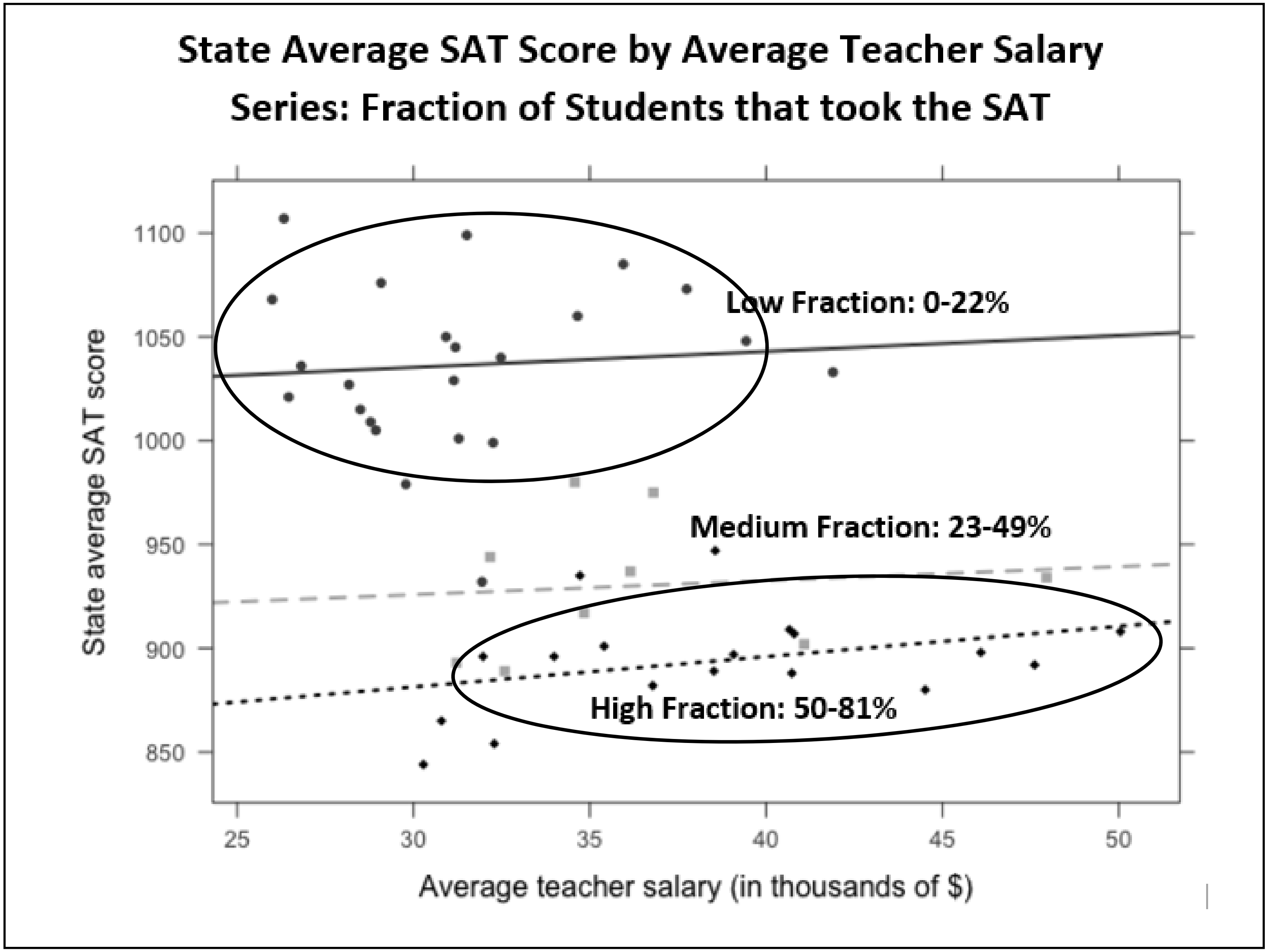


This mosaic plot doesn't work well for me.

V0 2016 IASE-2 7

Show Multivariable:
#2: XY Plot (2 factors)




State Average SAT Score by Average Teacher Salary
Series: Fraction of Students that took the SAT
Low Fraction: 0-22%
Medium Fraction: 23-49%
High Fraction: 50-81%
State average SAT score
1100
1050
1000
950
900
850
25
30
35
40
45
50
Average teacher salary (in thousands of $)

#2 Show Multivariable: Confounder is Too Complex

This method models separate series in that same XY plot. The confounder: percentage of students in the state that took the SAT.

- Consider the "low-fraction" states in the upper-left corner. Most students in the Middle states take the ACT – not the SAT. Only the best "middle" students take the SAT in applying to colleges on the East or West coast. In the "middle" teacher salaries are lower.
- Consider the "high fraction" states in the lower-right corner. Most students on the East and West coast take the SAT. These students include all students: best, middle and below-average so their average SAT is lower. On the coasts, teacher salaries are higher.

Controlling for the percentage taking the SAT changes the association between teacher salaries and average student scores.

#3 Show Multivariable Regression X-Y Output

Scottish Hill Races (Time in seconds)

Response variable is: Women's Record
R squared = 85.2% R squared (adjusted) = 84.9%
s = 1126 with 70-2 = 68 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	320.528	222.2	1.44	0.1537
Climb	1.755	0.088	19.8	< 0.0001

Assume: All modelling assumptions are satisfied
Assume: All slope coefficients are statistically significant.
http://www.scottishhillracing.co.uk/

#3 Show Multivariate: Regression X1-X2-Y Output

Scottish Hill Races (Time in seconds)

Response variable is: Women's Record
R squared = 97.5% R squared (adjusted) = 97.4%
s = 468.0 with 70 - 3 = 67 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	-497.656	102.8	-4.84	< 0.0001
Distance	387.628	21.45	18.1	< 0.0001
Climb	0.852	0.0621	13.7	< 0.0001

Controlling for Distance decreases Climb coefficient from 1.755 to 0.852; increases R^2 from 85% to 97%.

2016 GAISE Appendix B: Closing Thoughts (1)

"Multivariable thinking is critical to make sense of the observational data around us. This type of thinking might be introduced in stages":

1. Learn to identify observational studies
2. Why randomized assignment … improves things
3. Wary: cause-effect conclusions from observational data
4. Consider – *and explain* -- confounding factors
5. Simple approaches (stratification) to *show* confounding

http://www.amstat.org/education/gaise/collegeupdate/GAISE2016_DRAFT.pdf

V0 2016 IASE-2 13

2016 GAISE Appendix B Closing Thoughts (2)

"If students do not have exposure to simple tools for disentangling complex relationships, they may dismiss statistics as an old-school discipline only suitable for small sample inference of randomized studies."

"This report recommends that students be introduced to multivariable thinking, preferably early in the introductory course and not as an afterthought at the end of the course."

V0 2016 IASE-2 14

GAISE 2016 Deletions

Making Room: What might be voted off the island?
(or at least slimmed down)

- Probability theory.
- Constructing plots by hand.
- Basic statistics by hand.
- Drills with *z*-, *t*-, $\chi 2$, and F-tables.
- Advanced training on a statistics software program.

Five Other Methods for Presenting Confounding

A. Show confounding

1. Stratification using 2x2 averages tables
2. Stratification using 2x2 rate tables

B. Explain confounding

1. Mixture Displays
2. Wainer diagrams
3. Reverse-engineering rate tables

A1: Show Confounding: Stratified 2x2 Averages Table

At age 20, the average male-female weight difference is:

27 pounds [156 – 129] Average cells have grey fill.

Ave Weight	Ht=64"	Ht=70"
FEMALE	129	142
MALE		156

14 pounds [156-142] **after controlling for height**.

* www.cdc.gov/growthcharts/html_charts/bmiagerev.htm

V0 2016 IASE-2 17

A2: Show Confounding: Stratified 2x2 Rate Tables

Death Rates by Group

DIED	YOUNG	OLD	TOTAL
NON SMOKER	12%	86%	31%
SMOKER	18%	88%	24%
TOTAL	15%	86%	28%

Non-smokers are more likely to die than smokers

Within Young (and within Old), the reverse is true.

Problem with "Showing" Confounding

1. Do these visualizations "explain" confounding?
2. Can students use these devices to work problems with numerical answers?
3. Will any of this be on the final?

If all three answers are "No", teachers are unlikely to spend much time showing multivariable thinking.

Maybe during the last class before the final ☺

B1. Explain Confounding: Explicit Mixture Displays

	Year 1			Year 2		
	Number	Score		#	Score	
Disavantaged	10	80%		50	81%	⇧
Advantaged	90	90%		50	91%	⇧
TOTAL	100	89%		100	86%	⇩

After Year 1, other disadvantaged student switch to this teacher increasing their prevalence from 10% to 50%.

Teacher's scores: Better for each group; worse overall.

Explanation: "It's the mix"

B2. Explain Confounding: Wainer's Standardization

Wainer (2004) introduced a graphical technique that controlled for the influence of a binary confounder.

It requires minimal math and is visually intuitive.

My music and art majors find this graph easy to read. They can work problems with numerical answers.

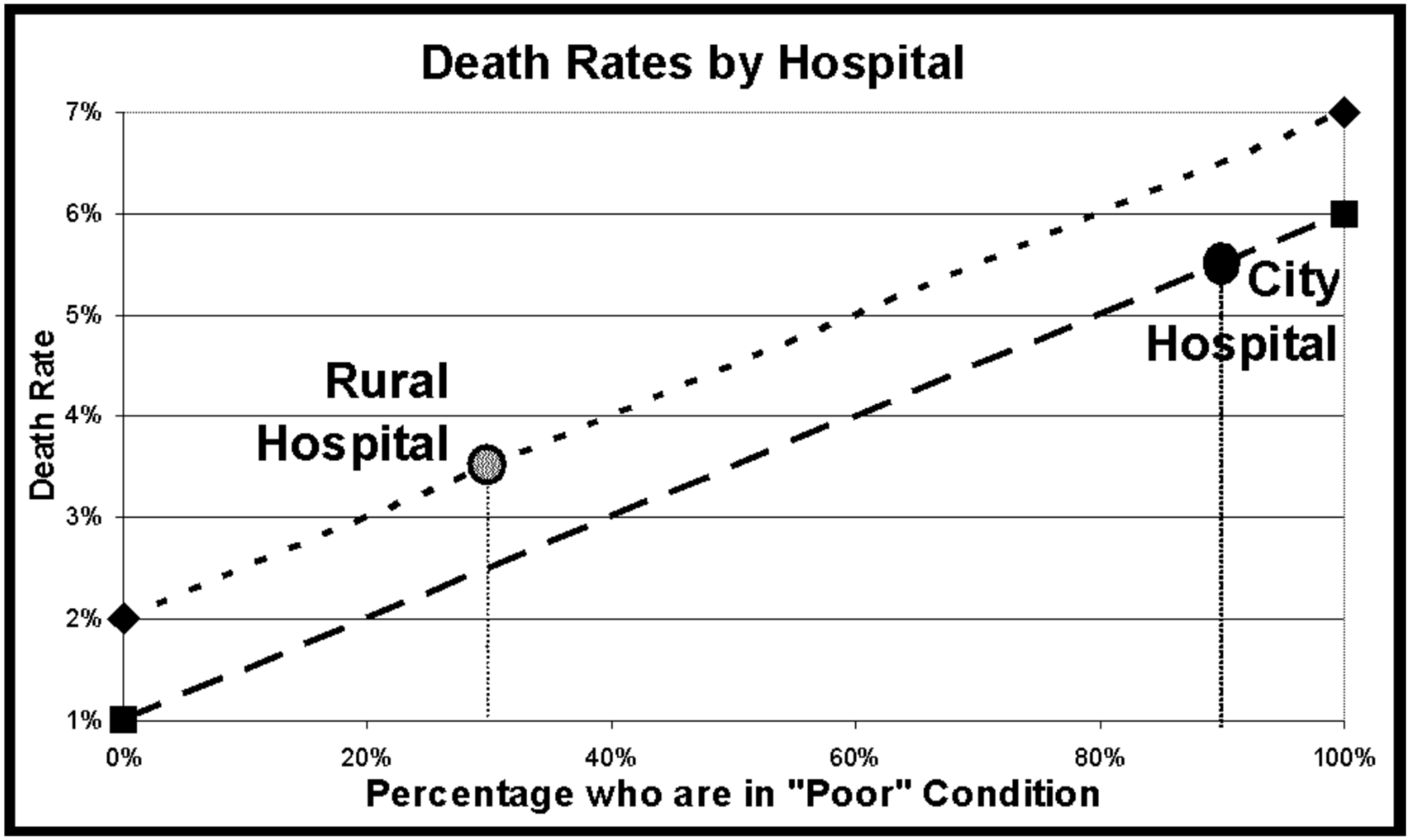
For the origin (1986) and details, see
> Tan (2012): www.statlit.org/pdf/2012-Tan-Simpsons-Paradox.pdf
> Schield (2006): www.statlit.org/pdf/2006SchieldSTATS.pdf.

V0

2016 IASE-2

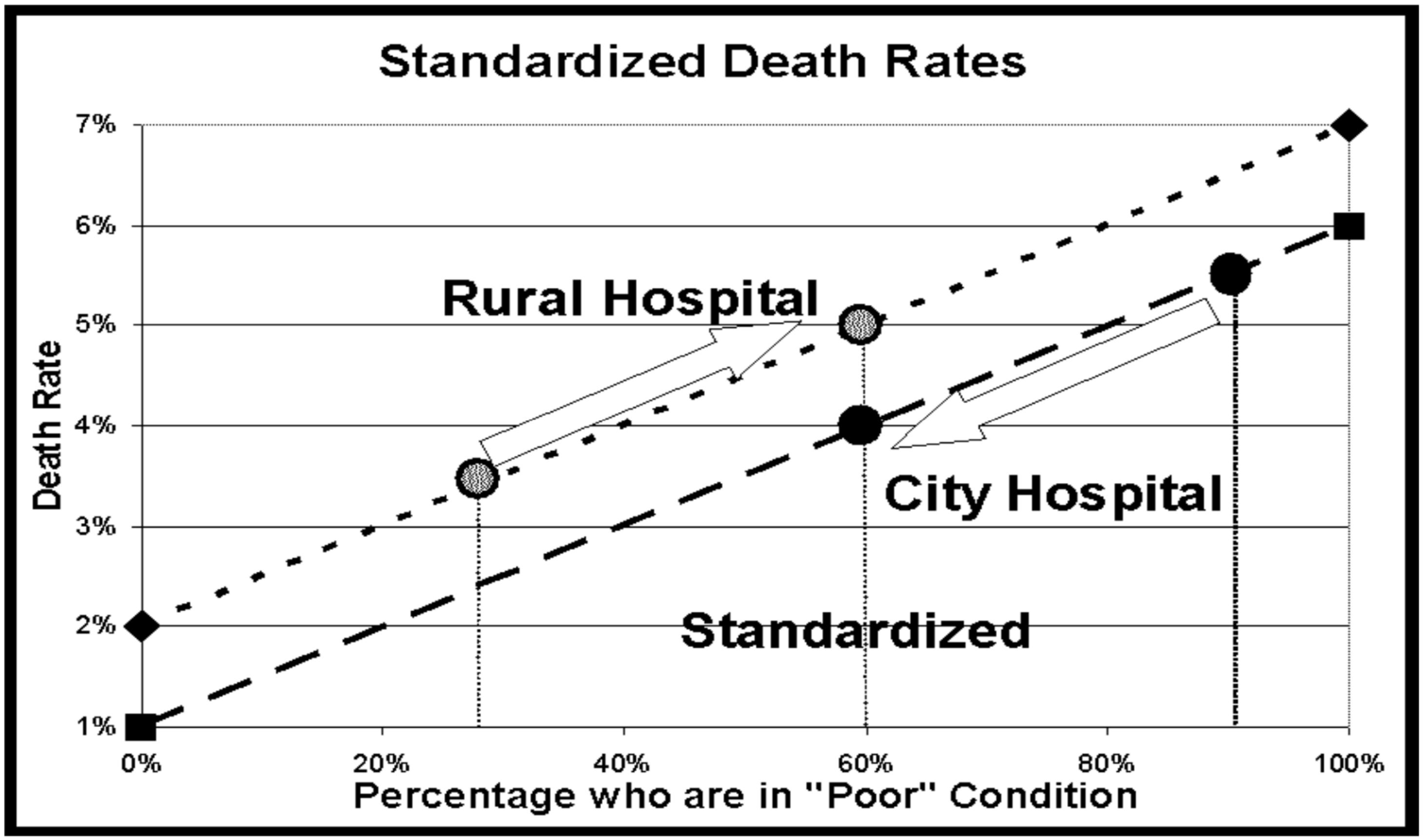
21

#B2: Wainer Diagrams
Simpson's Paradox: It's the Mix



V0 2016 IASE-2 22

Simpson's Paradox: It's the Mix
Standardize: Common Mixture



V0 2016 IASE-2 23

B3. Explain Confounding: Reverse-Engineer Rate Tables

DIED	YOUNG	OLD	TOTAL
NON SMOKER	12%	86%	31%
SMOKER	18%	88%	24%
TOTAL	15%	86%	28%

74% of top row are young; 90% of Row 2 are young.
82% of Row 3 are young; standardize top 2 with 82% young
Non-smoker standard death rate: 25% (0.82*12+0.18*86)
Smoker standardized death rate: 31% (0.82*18+0.18*88)
Standardized death rate for smokers > than for non-smokers

V0 2016 IASE 24

Why Statistical Educators Won't Teach Confounding

1. Students will have less trust in statistics if any confounder can reverse any association
2. Statisticians are not subject-matter experts
3. Emphasizes inductive/hypothetical thinking
4. Co-variation and sufficiency are math; confounding and causation are not.
5. "Association is not causation". K. Pearson: Causation is "a fetish amidst the inscrutable arcana of modern science"

"Less Trust" vs. Cornfield Conditions

1950s: Fisher said that the smoking-death (10X) association might be confounded by genetics (3X).

Cornfield proved that to nullify (or reverse) this association, the confounder must exceed 10X.

"Cornfield's minimum effect size is as important to observational studies as is the use of randomized assignment to experimental studies." Schield (1999)

Schield (1999) www.statlit.org/pdf/1999SchieldASA.pdf

V0 2016 IASE-2 26

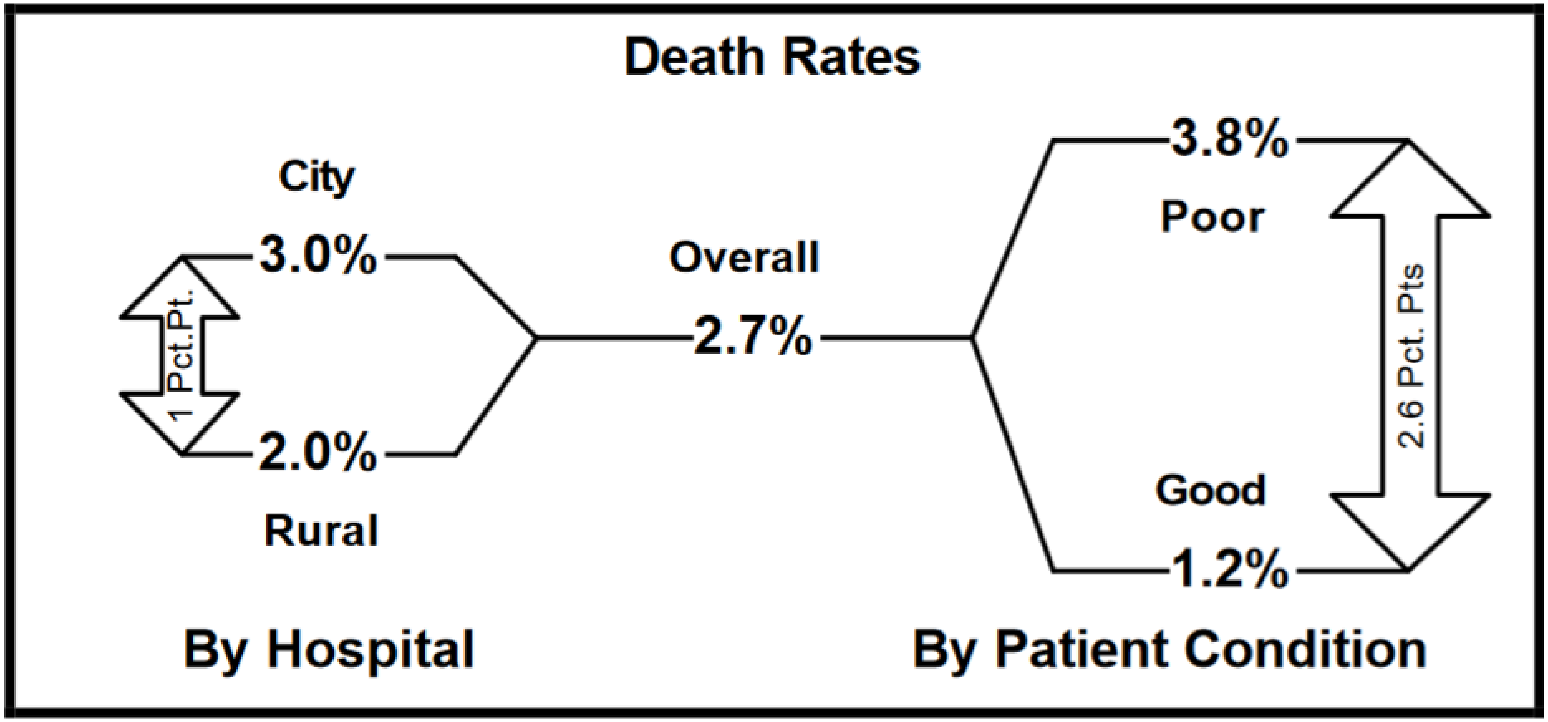
Stratification
Two-Way Half Tables

Patient Died	"Good"	"Poor"	TOTAL
City Hospital	1%	6%	5.5%
Rural Hospital	2%	7%	3.5%
TOTAL	1.5%	6.5%	

City patient is 2 pts more likely to die that a Rural patient.
Poor patient is 5 pts more likely to die than a Good patient.
Association with Outcome: Confounder > Predictor

V0 2016 IASE 27

Cornfield Condition for Nullification or Reversal



Schield (1999) based on realistic data

V0 2016 IASE 28

Cornfield Condition for Nullification or Reversal

An association is nullified or reversed ***only if***

- *confounder (patient condition) has a stronger association with the outcome (death) than does the predictor (hospital).*
- *predictor (hospital) has a stronger association with the confounder (patient condition) than with the outcome (death).*

Teaching Confounding

The bigger the effect size, the less likely a confounder can negate or reverse and observed association.

Effect Sizes:

- 10X: Smoking and death from lung cancer
- 1.3X: Second hand smoke and death

30

Confounded