



Simpson's Paradox, Confounding Variables and Insurance Ratemaking

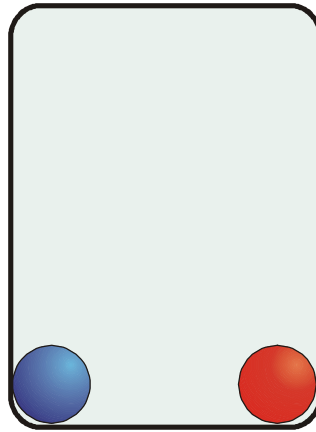
John Stenmark, FCAS, MAAA

Peter Wu, FCAS, ASA, MAAA

Presented at the 2004 Annual Meeting of the Casualty Actuarial Society

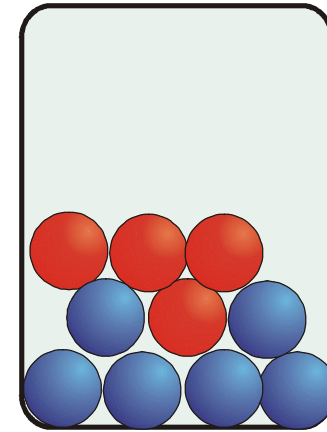
November 16, 2004

Simpson's Paradox



A

$$1/2=50\%$$

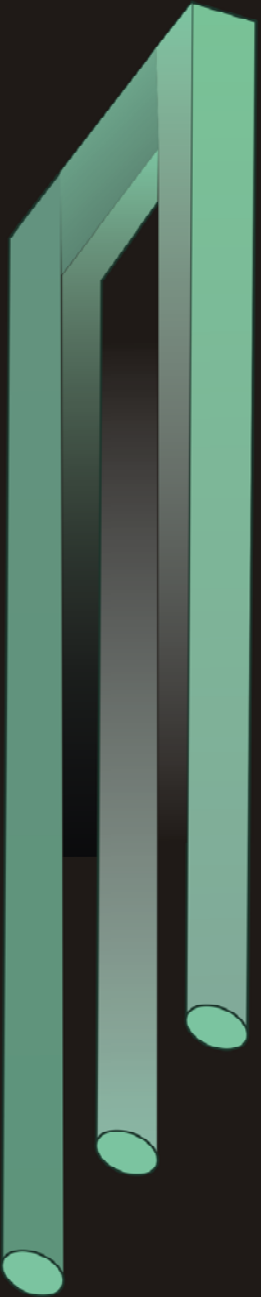


B

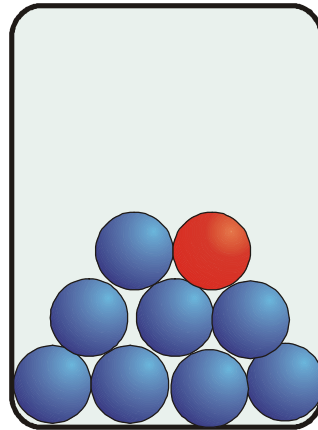
$$4/10=40\%$$

From which urn do you have the greater probability of drawing a red ball?

You have the greater probability of drawing a red ball from urn A

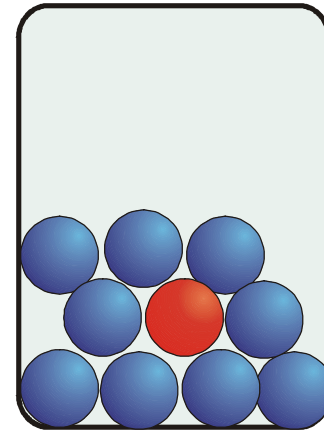


Simpson's Paradox



a

$$1/9=11\%$$

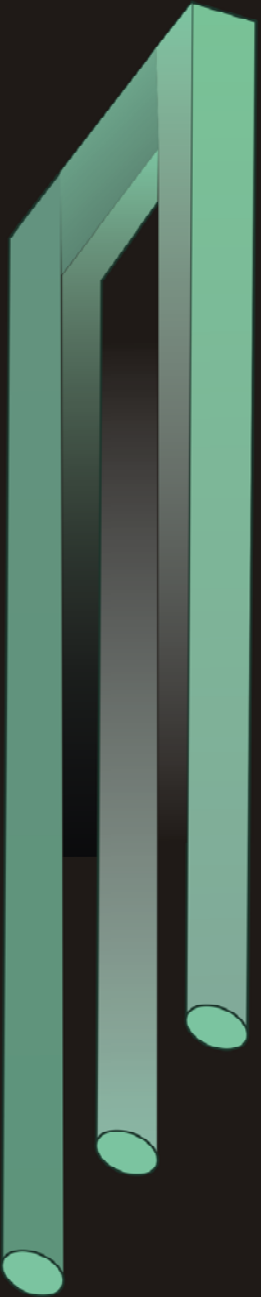


b

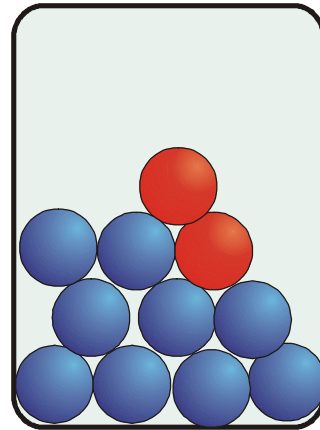
$$1/10=10\%$$

From which urn do you have the greater probability of drawing a red ball?

Again, you have the greater probability of drawing a red ball from urn a

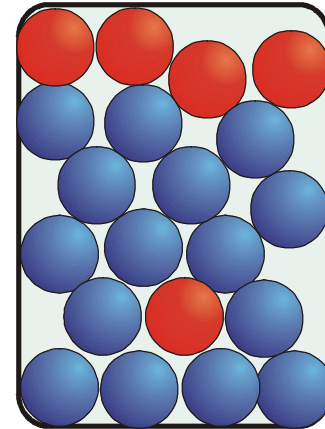


Simpson's Paradox



$a + A$

$$(1+1)/(2+9)=2/11=\underline{\underline{18\%}}$$



$b + B$

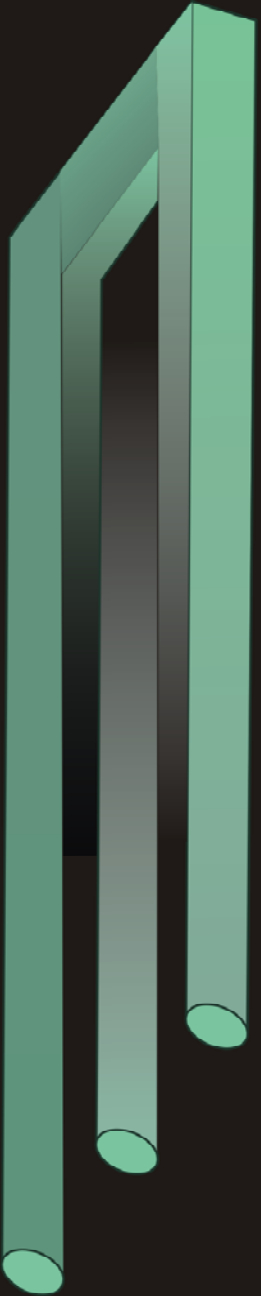
$$(4+1)/(10+10)=5/20=\underline{\underline{25\%}}$$

Let's pour urn a into urn A and urn b into urn B. Now from which urn do you have the greater probability of drawing a red ball?

This time you have a greater probability of drawing a red ball from urn b+B. Why?

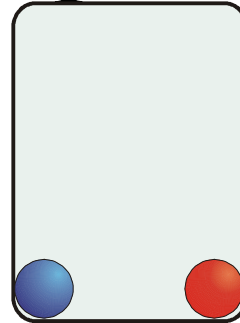
Overview

- Math behind Confounding and Simpson's Paradox
- Definition of a Confounding Variable
- Experimental Design
- Types of Confounding Variables
- Treatment of Confounding Variables
- Conclusions

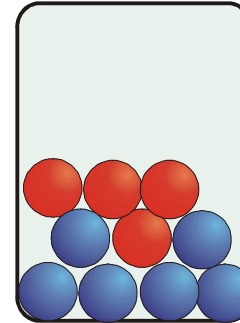


Simpson's Paradox

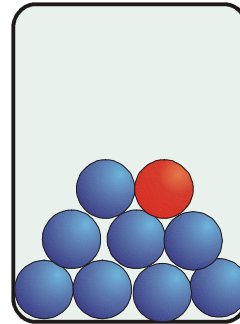
A
50%



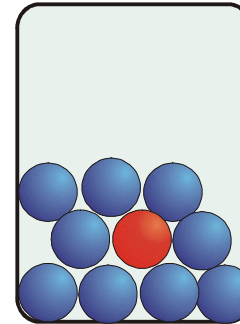
B
40%



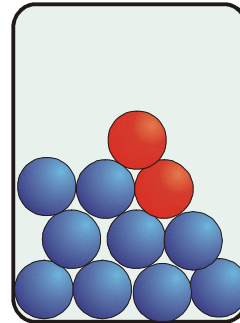
a
11%



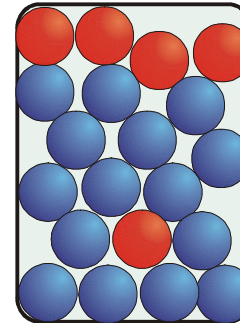
b
10%



$a + A$
18%






$b + B$
25%



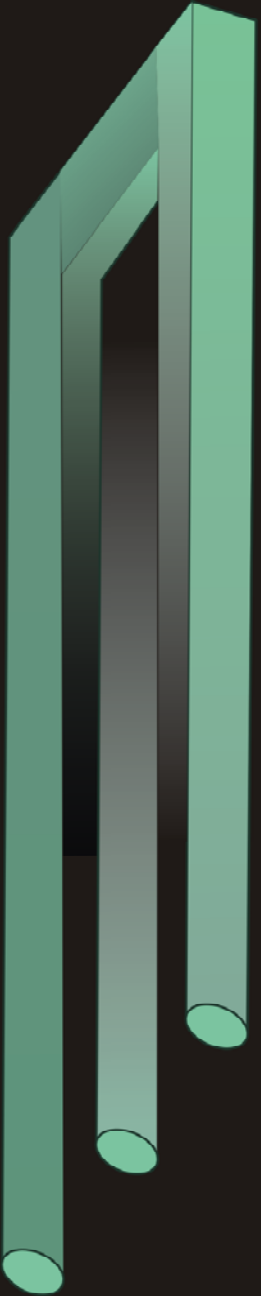
What does this have to do with insurance ratemaking?

Simpson's Paradox

and its relationship to Insurance Ratemaking

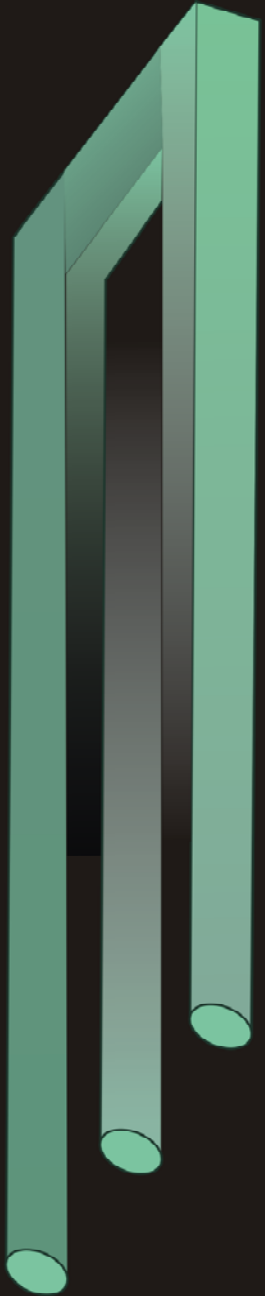
-  Assume that the red balls are insureds with claims.
-  Assume that the blue balls are insureds without claims.
-  Also assume that the balls in A and a are youthful insureds and the balls in B and b are adult insureds.

If the capital letters designate Massachusetts and the lower case letters designate Texas then the experience in each state separately indicates that the youthful drivers should be charged more than adults.



Simpson's Paradox

But when the balls
(experience) are combined
then that experience indicates
that the adults should be
charged more than youthful
drivers.



Simpson's Paradox

Without Good Student Discount					
	Exposures	%	Losses	Pure Premium	
Age 15-20	45	60%	4,500	100.00	
Age 21-25	99	99%	4,950	50.00	
Total	144		9,450	65.63	

With Good Student Discount					
	Exposures	%	Losses	Pure Premium	Relativity
Age 15-20	30	40%	2,400	80.00	-20%
Age 21-25	1	1%	40	40.00	-20%
Total	31		2,440	78.71	20%

- Each group of youthful operators alone generates a 20% discount.
- Together they seem to justify a 20% surcharge.

Simpson's Paradox

Consider eight integers: A, B, C, D, a, b, c, d .

Now suppose that $\frac{a}{A} > \frac{b}{B}$ and $\frac{c}{C} > \frac{d}{D}$

Is it necessarily true that $\frac{a+c}{A+C} > \frac{b+d}{B+D}$?

It should be obvious that the answer is: ***No, not necessarily!***



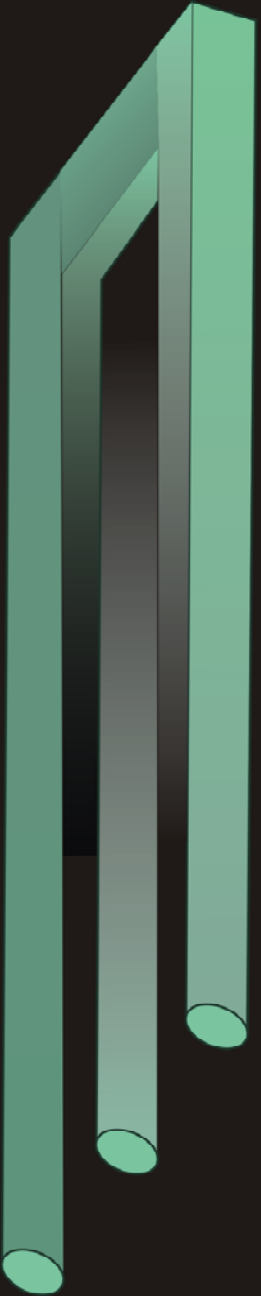
Simpson's Paradox

- The paradox occurs when a relationship or association between two variables reverses when a third factor, called a confounding variable, is introduced.
- The paradox also occurs when a relationship/association reverses when the data is aggregated over a confounding variable.



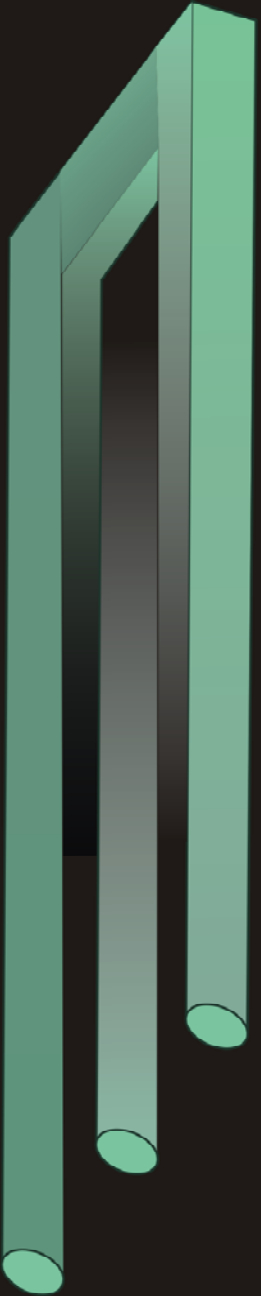
Confounding Variables

A variable can confound the results of a statistical analysis only if it is related (non-independent) to both the dependent variable and at least one of the other (independent) variables in the analysis.



Confounding Variables

More specifically, a variable can confound the results of an insurance rate structure analysis only if it is related (non-independent) to both the experience measure (loss ratio, pure premium, etc.) and at least one of the other rating variables in the analysis.



Confounding Variables

Again consider the eight integers: A, B, C, D, a, b, c, d.

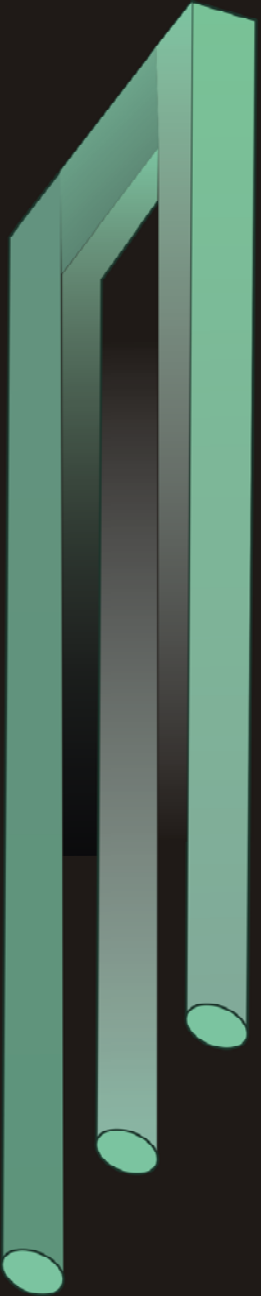
Now suppose that $\frac{a}{A} - \frac{b}{B} = K = \frac{c}{C} - \frac{d}{D}$

Is it necessarily true that $\frac{a+c}{A+C} - \frac{b+d}{B+D} = K$?

It should be obvious that the answer is, again: *No, not necessarily!*

Experimental Design

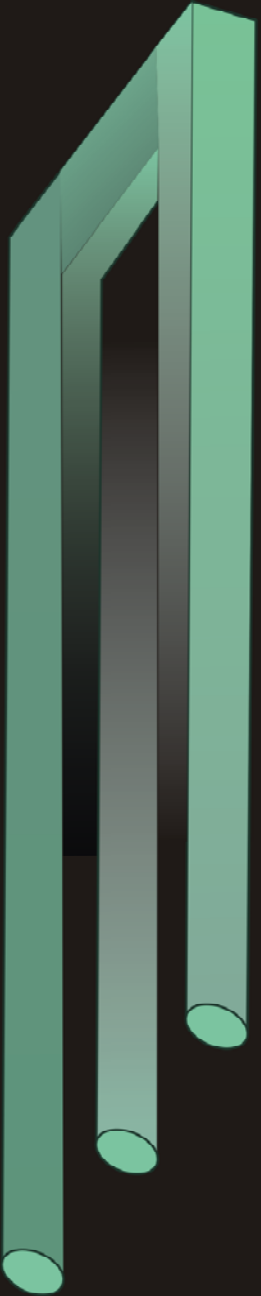
- An important statistical subject in designing researches and studies:
 - Determining influential variables
 - Minimizing variability of experimental results
 - Maximizing efficiency and reducing cost
 - Minimizing influences of uncontrollable confounding variables
- Confounding and Simpson's paradox is a great concern in designing research experiments



Experimental Design

“Ideal designs” for research experiments:

- Balanced design: equal number of collected “data points” for every combination of predictive variables
- Proportional design: distribution of collected “data points” is proportional (uniform) across predictive variables – I.e. predictive variables’ distributions are independent to each other



Experimental Design – A Chemistry Experiment Example

Balanced Design – Number of Observations

Concentration	Temperature		
	32°F	122°F	212°F
1%	10	10	10
2%	10	10	10

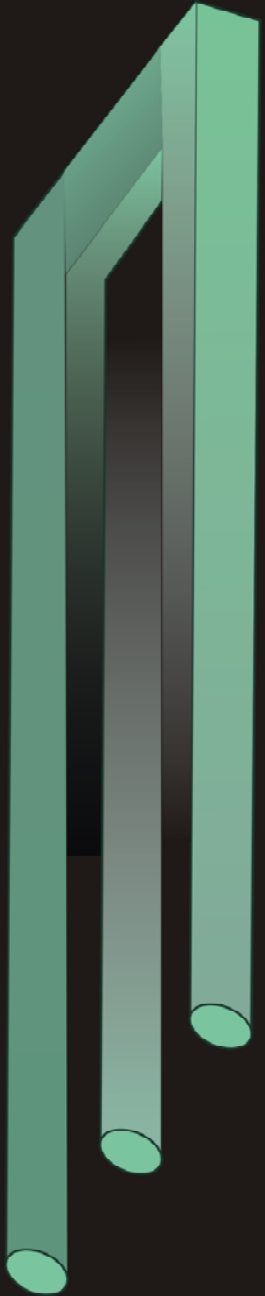
Proportional Design – Number of Observations

Concentration	Temperature		
	32°F	122°F	212°F
1%	2	5	10
2%	4	10	20



Experimental Design

Important Principle: *If there is independence between the potential confounding variable and the variable under study, or if the study is balanced or proportionally distributed, then there is no confounding.*



Confounding Variables

Again consider the eight integers: A, B, C, D, a, b, c, d.

Now suppose that $\frac{a}{A} - \frac{b}{B} = K = \frac{c}{C} - \frac{d}{D}$

Is it necessarily true that $\frac{a+c}{A+C} - \frac{b+d}{B+D} = K$?

Is there some property or properties of A, B, C, D, a, b, c and d for which the above equation is true?



Confounding Variables

Important Principle: *If there is independence between the potential confounding variable and the variable under study, or if the study is balanced or proportionally distributed, then there is no confounding.*

“independence between the potential confounding variable and the variable under study”

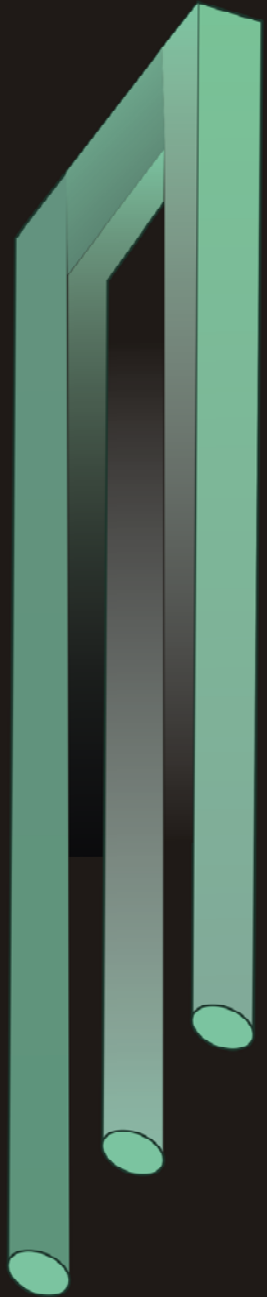
$$\frac{a}{A} = \frac{c}{C} \text{ and } \frac{b}{B} = \frac{d}{D}$$

“the study is proportionally distributed”

$$\frac{A}{B} = \frac{C}{D}$$

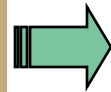
“the study is balanced ”

$$A = C \text{ and } B = D$$



Confounding Variables

"the study is proportionally distributed"



$$\frac{A}{C} = \frac{B}{D} = K'$$

*Then $A = CK'$
and $B = DK'$*

Remember that $\frac{a}{A} - \frac{b}{B} = K = \frac{c}{C} - \frac{d}{D}$

$$\frac{a+c}{A+C} - \frac{b+d}{B+D} = \frac{a+c}{CK'+C} - \frac{b+d}{DK'+D} = \frac{1}{K'+1} \left(\frac{a+c}{C} - \frac{b+d}{D} \right)$$

$$= \frac{1}{K'+1} \left(\frac{a}{C} + \frac{c}{C} - \frac{b}{D} - \frac{d}{D} \right) = \frac{1}{K'+1} \left(\frac{a}{C} - \frac{b}{D} + K \right) = \frac{1}{K'+1} \left(\frac{a}{\frac{C}{K'}} - \frac{b}{\frac{D}{K'}} + K \right)$$

$$= \frac{1}{K'+1} \left[K' \left(\frac{a}{A} - \frac{b}{B} \right) + K \right] = \frac{1}{K'+1} (K'K + K) = \frac{K'+1}{K'+1} K = K$$

Confounding Variables

Specific Comments to Insurance

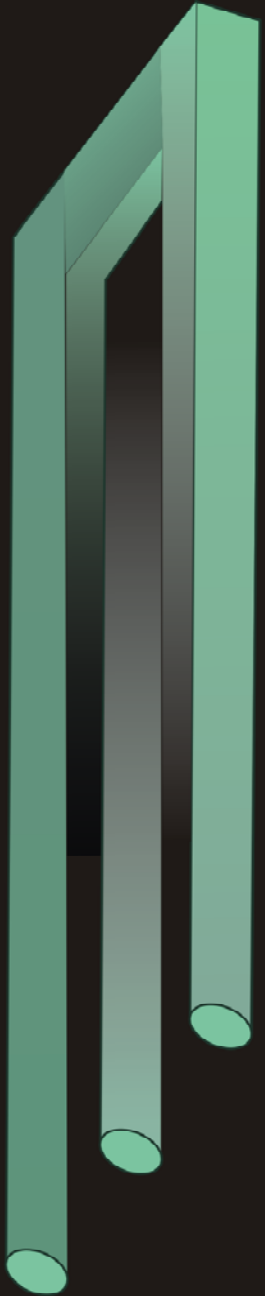
Insurance ratemaking differs from most statistical studies in a number of ways:

1. It is generally not possible to design the makeup of groups of insureds so that classifications are balanced.
2. Generally there are far more values for each variable and probably more variables in insurance than in research analysis.
3. In most statistical studies the objective is to accept or reject a hypothesis. The primary concern in insurance ratemaking is to properly calculate a rate, which requires a continuous rather than binary output.



Types of Confounding Variables

- Stratification Confounding Variable
- Aggregation Confounding Variable
- Lurking Confounding Variable



Stratification Confounding Variable

- Data can be stratified by the numerous elements collected
- Often may be rating elements or potential rating elements
- As the data is stratified each cell contains a smaller sample encouraging the aggregation of data from more sources which leads to more confounding...

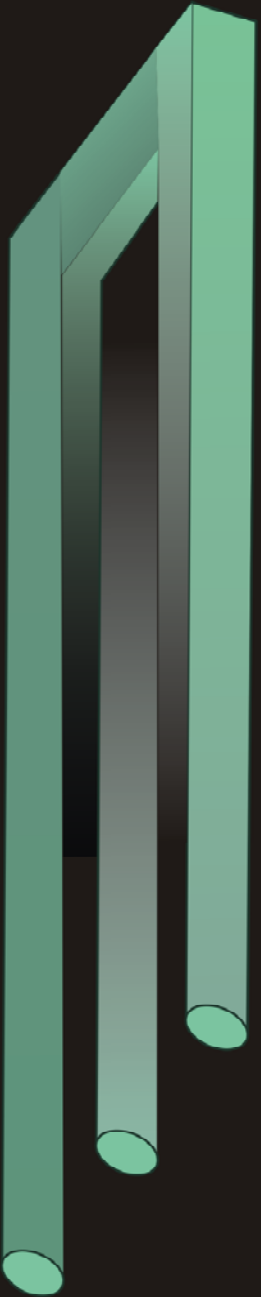


Aggregation Confounding Variable

“It’s a well accepted rule of thumb that the larger the data set, the more reliable the conclusions drawn. Simpson’s paradox, however, slams a hammer down on the rule and the result is a good deal worse than a sore thumb.

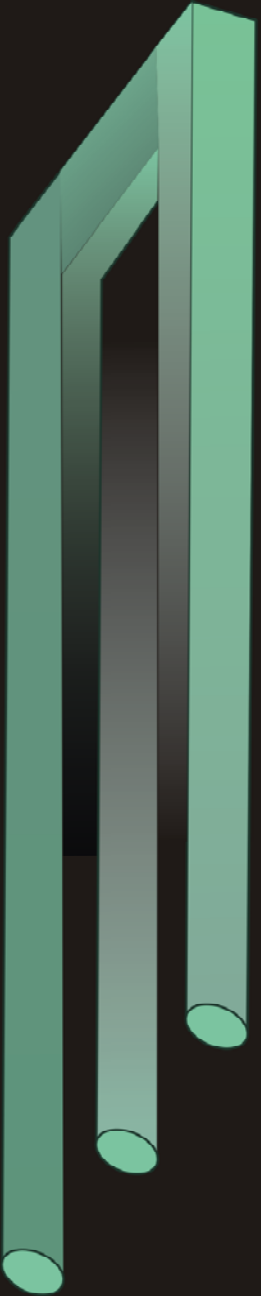
*Unfortunately Simpson’s paradox demonstrates that a great deal of care has to be taken when combining small data sets into a large one. Sometimes conclusions from the large data set are exactly the opposite of conclusion from the smaller sets. Unfortunately, the conclusions from the large set are also usually wrong.” **

* From “Simpson’s Paradox - When Big Data Sets Go Bad”



Aggregation Confounding Variable

- In order to increase the volume of data available for an analysis data from more than one year and/or state is aggregated.
- Since different years or states may have different distributions and different profitability a confounding effect can be created
- Consider Model Year ratemaking, for example.



Lurking Confounding Variable

- A variable that has not yet been discovered to be a confounding variable
- May be in company database, external data source or not exist as measured data at all
- Use of Credit Score illustrates the discovery of a Confounding Variable



Lurking Confounding Variable

There are two issues relative to the discussion of confounding in previously unused rating variables, such as credit:

- Prior to its use as a rating variable, the failure to segment insureds according to any credit measure may have caused confounding of those rating variables actually in use.
- Once credit score has been established as a rating variable proper methods must be undertaken to prevent the continued confounding of the class rates through the use of one of the treatments described in the next section.



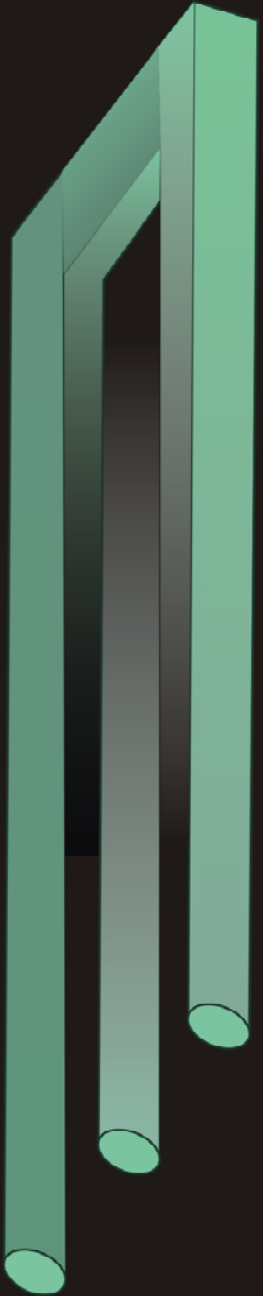
Treatment of Confounding Variables

- No Treatment
- Controlling Confounding through **Experimental Design**
- Controlling Confounding through **Multivariate Analysis**
- Controlling Confounding through the Use of **Meta-analysis**
- Controlling Confounding through the Use of **Scaling Factors**

Controlling Confounding through the use of

Scaling Factors

Is there a way that data from several experience years and several states can be aggregated to increase data volume without possibly confounding the results of the study and without the necessity of inclusion of the confounding variable in the analysis?



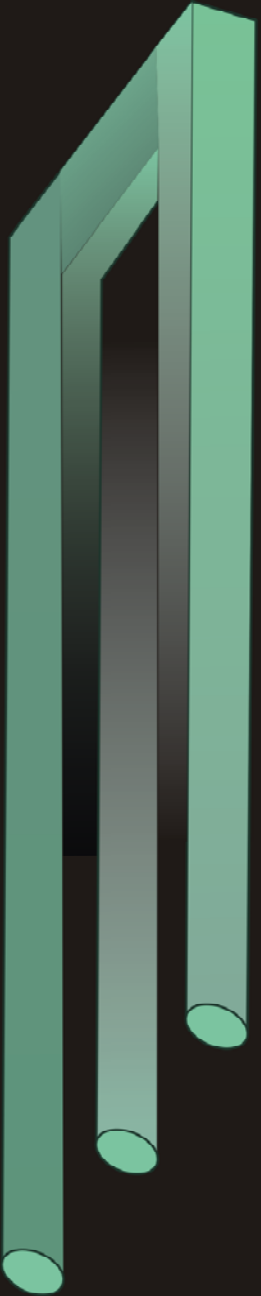
Controlling Confounding through the use of

Scaling Factors

As stated previously, there are two conditions necessary for a variable to confound the results of an analysis:

- 1. There must be a relationship between that variable and the experience variable.*
- 2. There must be a relationship between that variable and at least one of the rating variables under analysis.*

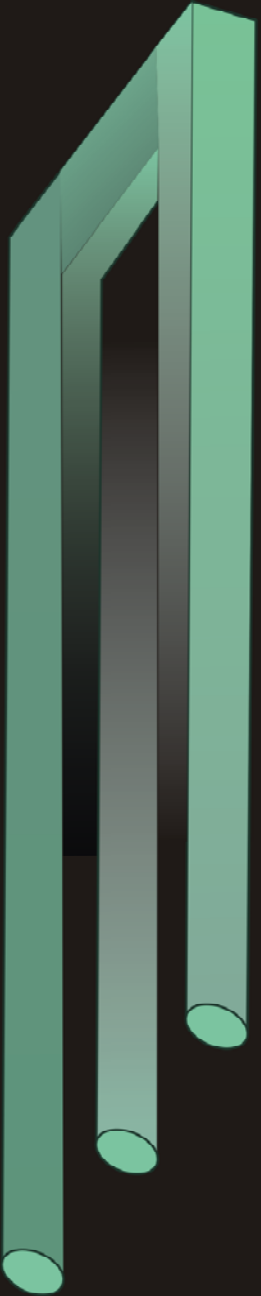
If either of those two conditions is not met then there is no confounding of the results.



Controlling Confounding through the use of

Scaling Factors

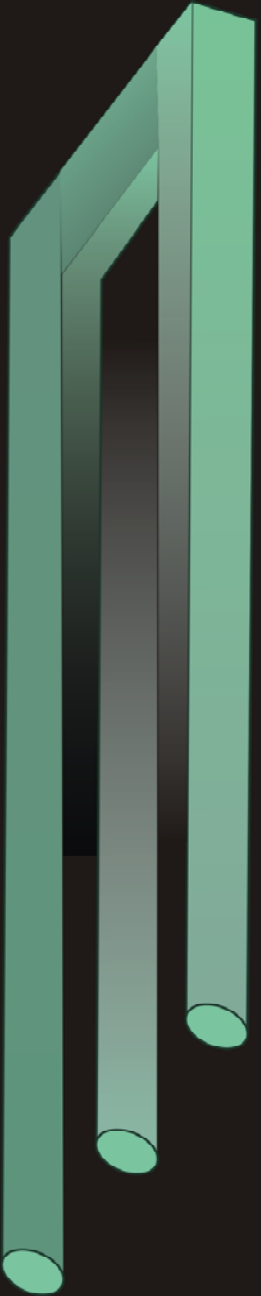
- If both conditions are met can the data be modified so that one of the conditions is no longer met, eliminating the confounding?
- Must be done in such a manner that the important underlying relationships in the data are not disturbed.
- In the following slides, we will show the scaling factors approach using a class plan analysis example with two potential confounding variables – states and years.



Controlling Confounding through the use of

Scaling Factors

- First a loss model is introduced.
- The model is specific to the ratemaking methodology used as well as whether base class or overall average class is used.
- The model is then used to determine if a scaling factor candidate does, in fact, eliminate the bias.



Simpson's Paradox, Confounding Variables and Insurance Ratemaking

Controlling Confounding through the use of

Scaling Factors

The Loss Experience Model

This model is introduced to symbolically represent the bias in the determination of class factors when multiple states and years are used.

First the calculated (base class) factor (for the modified loss ratio method) is represented to the right.

Then the bias is represented below.

$$g_i = \frac{\sum_y \sum_s e_{iys} r_{ys} B_{ys} f_i}{\sum_y \sum_s e_{iys} B_{ys} c_b} \div \frac{\sum_y \sum_s e_{bys} r_{ys} B_{ys} f_b}{\sum_y \sum_s e_{bys} B_{ys} c_b}$$

The bias in the method

$$= \frac{g_i}{f_i} - 1 = \frac{\sum_y \sum_s e_{iys} r_{ys} B_{ys}}{\sum_y \sum_s e_{iys} B_{ys} c_b} \cdot \frac{\sum_y \sum_s e_{bys} B_{ys}}{\sum_y \sum_s e_{bys} r_{ys} B_{ys}} - 1$$

Controlling Confounding through the use of

Scaling Factors

The Loss Experience Model

The model can be modified to represent calculations based on an overall average modified loss ratio.

The calculated factor (for the modified loss ratio method) is represented to the right.

Then the bias is represented below.

$$g_i^O = \frac{\sum_y \sum_s e_{iys} r_{ys} B_{ys} f_i}{\sum_y \sum_s e_{iys} B_{ys} c_b} = \frac{\sum_i \sum_y \sum_s e_{iys} r_{ys} B_{ys} f_i}{\sum_i \sum_y \sum_s e_{iys} B_{ys} c_b}$$

The bias in the method

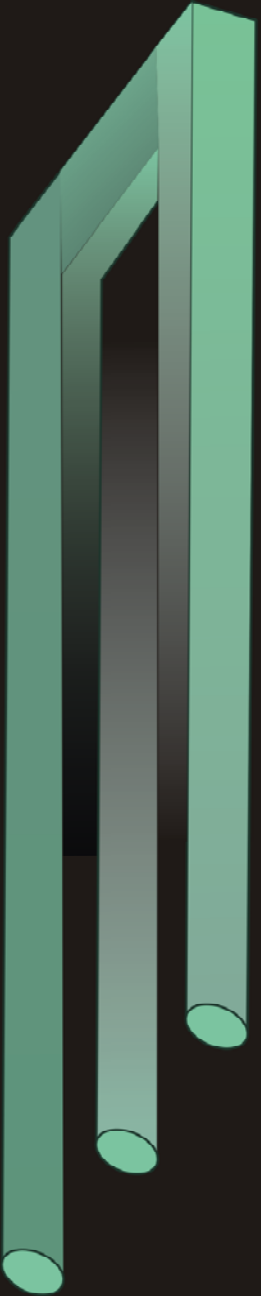
$$= \frac{g_i^O}{f_i^O} - 1 = \frac{1}{f_i^O} \frac{\sum_y \sum_s e_{iys} r_{ys} B_{ys} f_i}{\sum_y \sum_s e_{iys} B_{ys} c_b} \cdot \frac{\sum_i \sum_y \sum_s e_{iys} B_{ys} c_b}{\sum_i \sum_y \sum_s e_{iys} r_{ys} B_{ys} f_i} - 1$$



Controlling Confounding through the use of

Scaling Factors

- There are four scaling factors because ratemaking can use a base class or a statewide average as the base and because either the first or second condition can be addressed.
- Only one of the four types is used in preparing the data for an analysis.
- The factors may be the same for all classes and applied only to losses or they may be different for each class and applied to premiums and losses.
- In either case the indicated rate for each class for any confounding variable remains unaltered.



Scaling Factor 1

First Special Scaling Factor

$$s_{ys} = \frac{1}{r_{ys}}$$

- The reciprocal of the base class loss ratio.
- Applied to the losses only for a year and state.
- Eliminates the non-independence of the confounding variables (year and state) and the loss ratio.

Scaling Factor 2

Second Special Scaling Factor

$$S_{iys} = \frac{\sum_y \sum_s e_{iys}}{\sum_y \sum_s e_{bys}} \cdot \frac{e_{bys}}{e_{iys}}$$

- Uses the exposures of each class, the base class, all classes combined and all base classes.
- Different for every class, year and state
- Applied to the premiums and losses for each class.
- Eliminates the non-independence of the confounding variables (year and state) and class.

Scaling Factor 3

First Generalized Scaling Factor

$$S_{ys} = \frac{1}{r_{ys}^O}$$

- The reciprocal of the overall average class loss ratio.
- Applied to the losses only for a year and state.
- Eliminates the non-independence of the confounding variables (year and state) and the loss ratio.

Scaling Factor 4

Second Generalized Scaling Factor

$$S_{iys} = \frac{\sum_y \sum_s e_{iys}}{\sum_i \sum_y \sum_s e_{iys}} \cdot \frac{\sum_i e_{iys}}{e_{iys}}$$

- Uses the exposures of each class in year and state, class for all years and states combined, all classes, years and states combined and all classes combined for each year and state.
- Different factor for each class, year and state
- Applied to the premiums and losses for each class.
- Eliminates the non-independence of the confounding variables (year and state) and class.

Scaling Factors

Without Good Student Discount						
	Exposures	Scaling Factor	Scaled Exposures	Scaled Losses	Pure Premium	
Age 15-20	45	1.3714	61.71	6,171	100.00	
Age 21-25	99	0.8312	82.29	4,114	50.00	
Total	144		144.00	10,286	71.43	
With Good Student Discount						
	Exposures	Scaling Factor	Scaled Exposures	Scaled Losses	Pure Premium	Relativity
Age 15-20	30	0.4429	13.29	1,063	80.00	-20.0%
Age 21-25	1	17.7143	17.71	709	40.00	-20.0%
Total	31		31.00	1,771	57.14	-20.0%

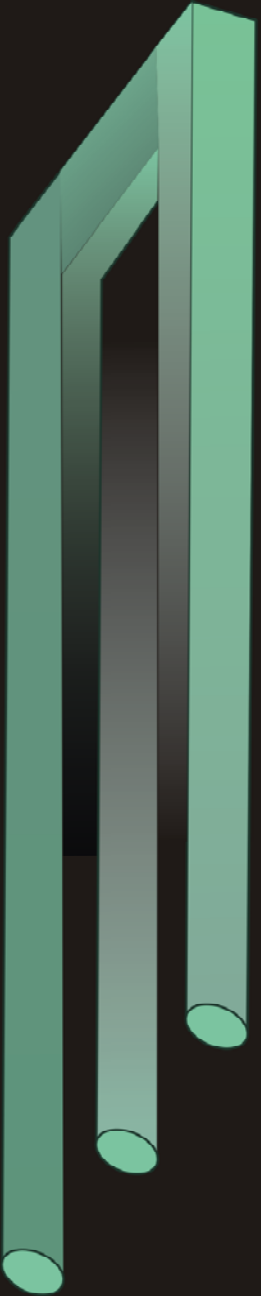
$$s_{ij} = \frac{\sum_i e_{ij}}{\sum_i \sum_j e_{ij}} \cdot \frac{\sum_j e_{ij}}{e_{ij}}$$

$$s_{1,1} = \frac{(45 + 30)}{(31 + 144)} \cdot \frac{144}{45} = 1.3714$$

- ...and the confounding effect is removed.

Conclusion

- Ratemaking precision is often compromised when rating variables are confounded by other variables
- Confounding variables can be addressed through the use of multivariate rating
- An alternative approach is to scale loss experience to eliminate the confounding effect





Simpson's Paradox, Confounding Variables and Insurance Ratemaking

John Stenmark and Peter Wu

Presented at the 2004 Annual Meeting of the Casualty Actuarial Society

November 15, 2004