A Bayes view on Simpson's paradox

2016-12-02
Bayes WG meeting, Mainz

Gerhard Nehmiz

Boehringer Ingelheim

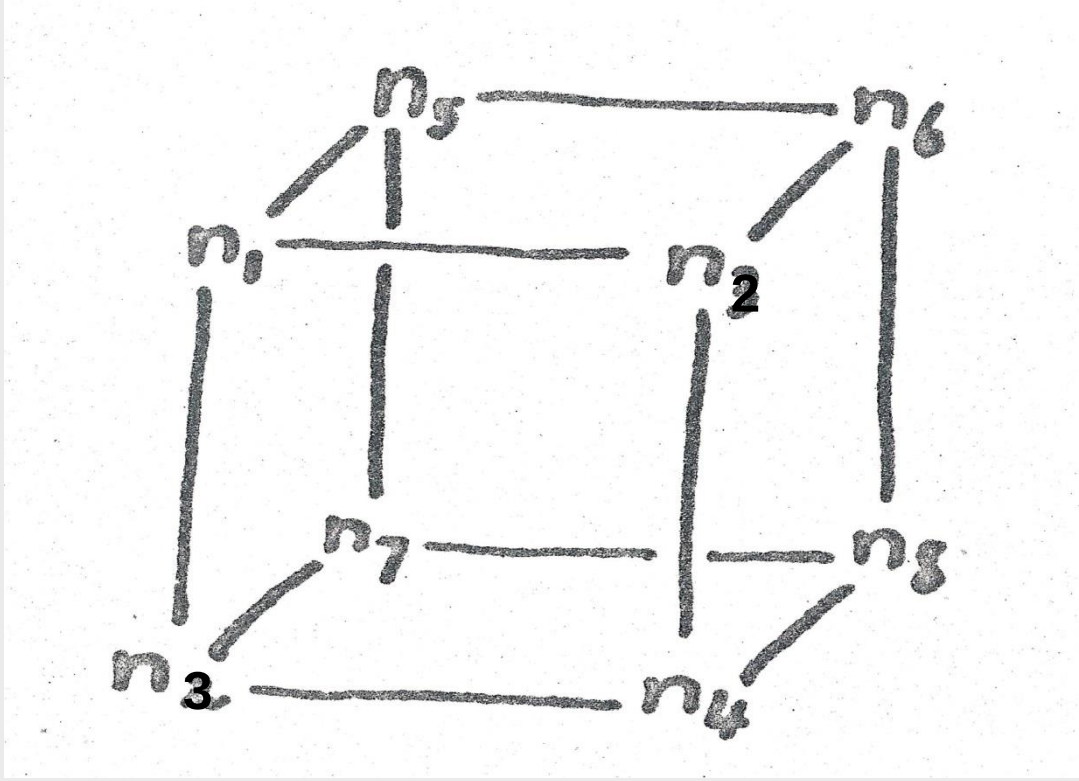
Overview

Boehringer Ingelheim

(1) Introduction
   (a) The nature of the problem
   (b) A basic example
(2) The prior probability for the Simpson phenomenon in the multinomial model
(3) The Bayes factor for presence or absence of the Simpson phenomenon
(4) Representation through a Directed Acyclic Graph (DAG)
(5) The meta-analysis example
(6) The continuity-correction example
(7) Discussion, outlook
(8) Literature

(1) Introduction

(a) The nature of the problem



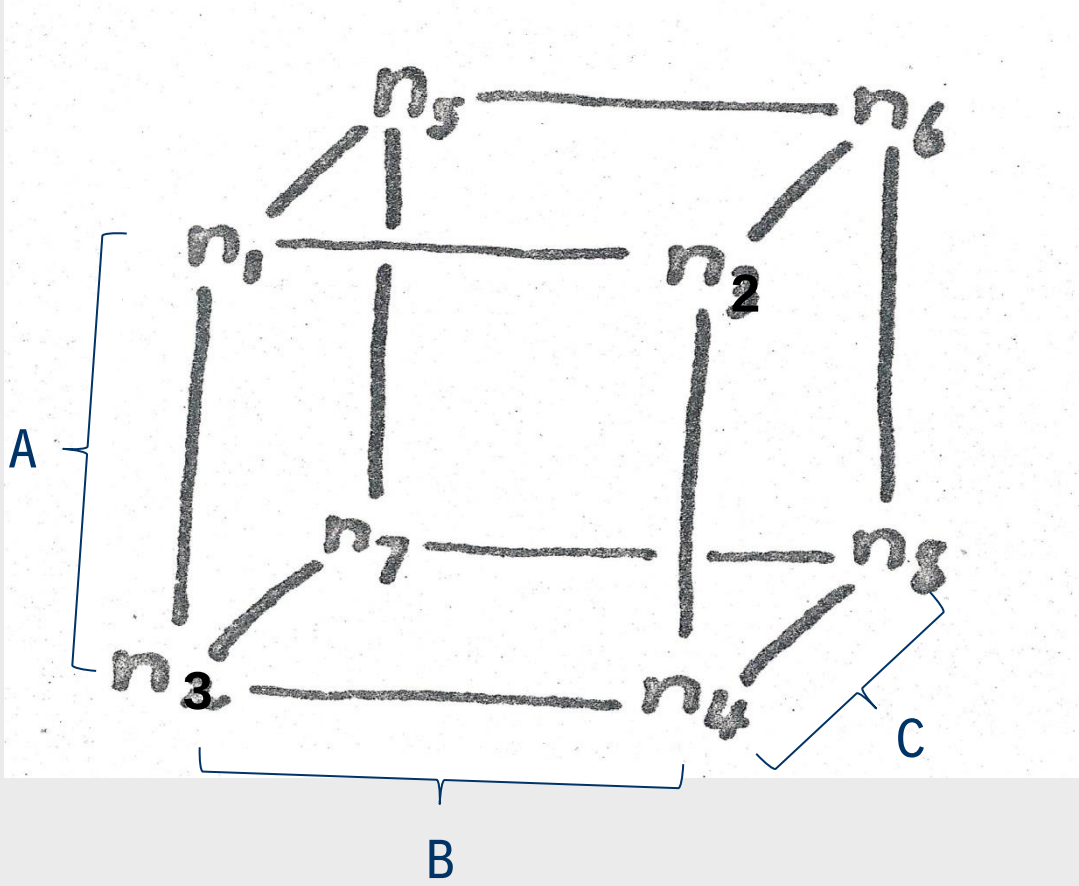
A 2x2xK frequency table. Here: K=2.

Note the re-numbering, it has no consequences for Bartlett's calculations as they are all symmetrical w.r.t. n_2 and n_3, but it is necessary for the symmetry (and also consistent with Bartlett's other drawing in the same article)

Bartlett, J.R.S.S.Suppl. 1935; Pavlides/Perlman, Am.Stat. 2009

Boehringer Ingelheim

(a) The nature of the problem



Samuels, J.A.S.A. 1993

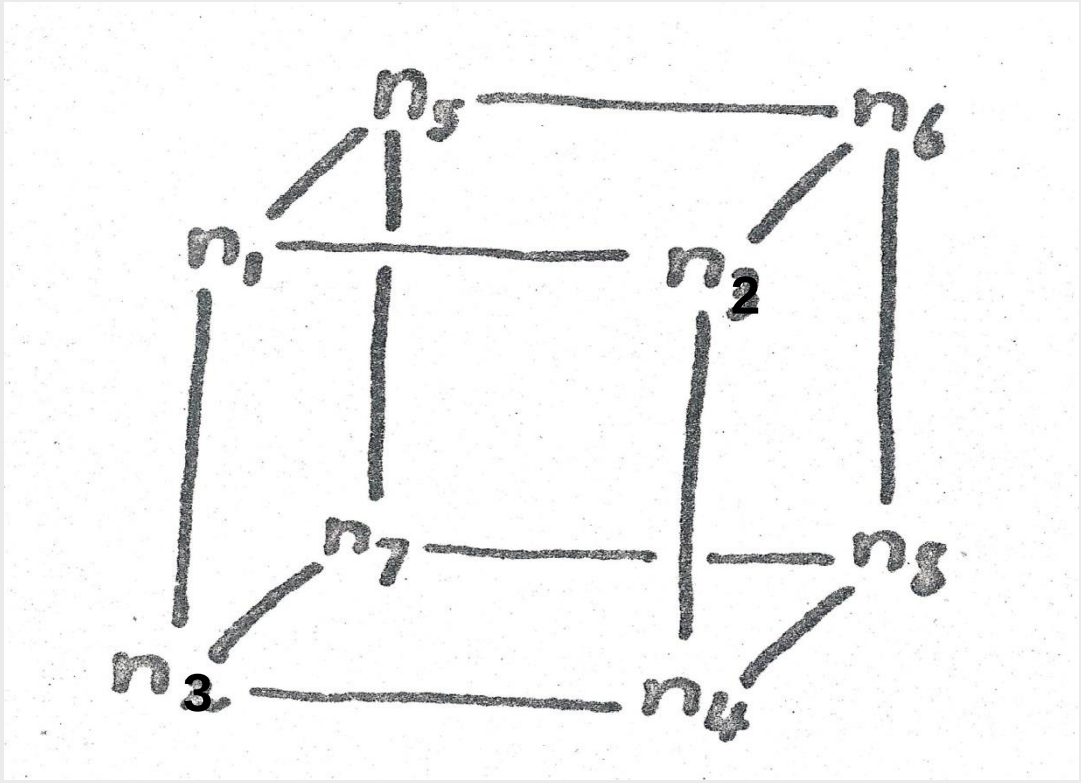
3 classifications.

Simpson‘s paradox is present if the association between A and B is in one direction (e.g. positive) conditionally for all values of C, but reversed (e.g. negative) when considered marginally over C.

C is a special type of confounder.

(1) Introduction

(a) The nature of the problem



A 2x2x2 frequency table.

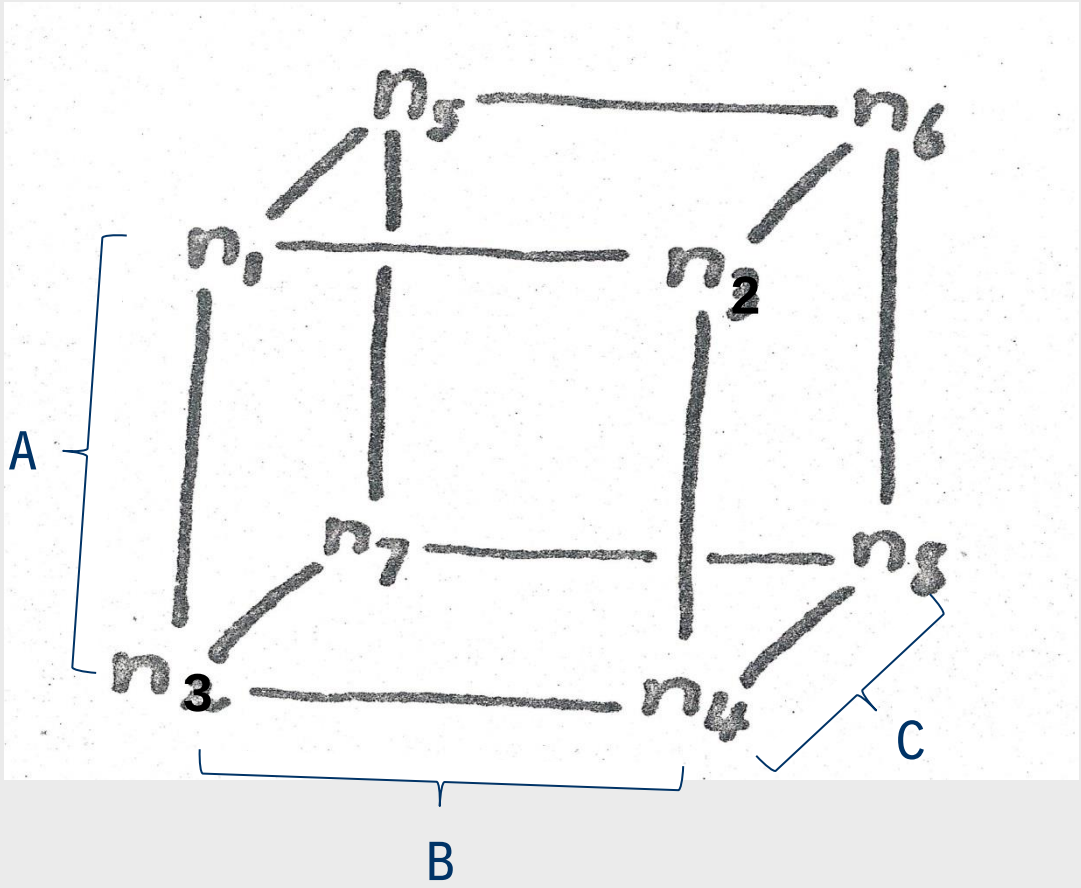
3 probability models for $n_{1..8}$:

- Multinomial for all 8 corners (i.e. arbitrary p_i's that sum up to 1)
- 4 x binomial: only p_1, p_2, p_5 and p_6 free, with fixed column sums (i.e. 2 independent variables and 1 dependent variable)
- conditional on fixed column and row sums in each layer

Good/Mittal, Ann.Stat. 1987

Boehringer Ingelheim

(b) A basic example



Julious/Mullee, B.M.J. 1994

Real examples are rare.

Yule 1903, Simpson 1951, Kendall/Stuart 1979, Chuang-Stein/ Beltangady 2011 are artificial.

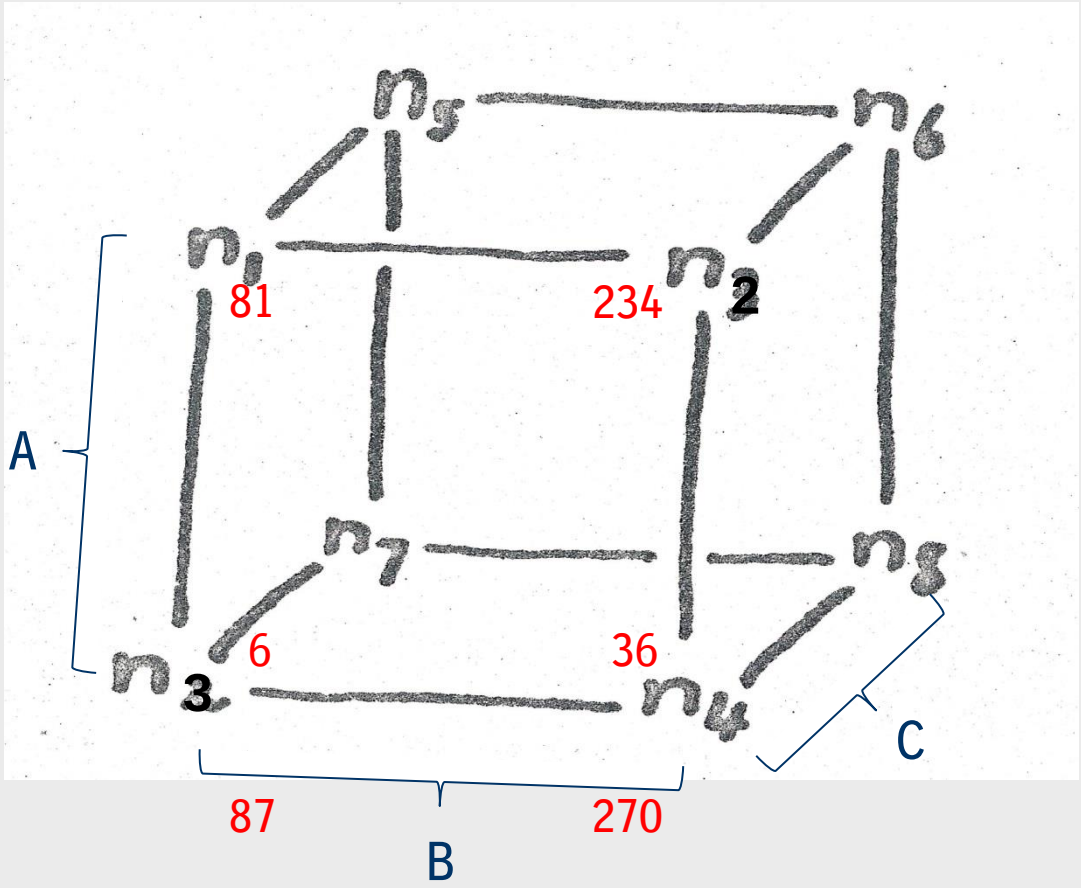
Julious/Mullee 1994: Kidney surgery.

A := success: yes/no,
B := type: open/percutaneous,
C := stone size class: small/large

(binomial model)

Boehringer Ingelheim

(b) A basic example



Julious/Mullee, B.M.J. 1994

Real examples are rare.

Yule 1903, Simpson 1951, Kendall/Stuart 1979, Chuang-Stein/ Beltangady 2011 are artificial.

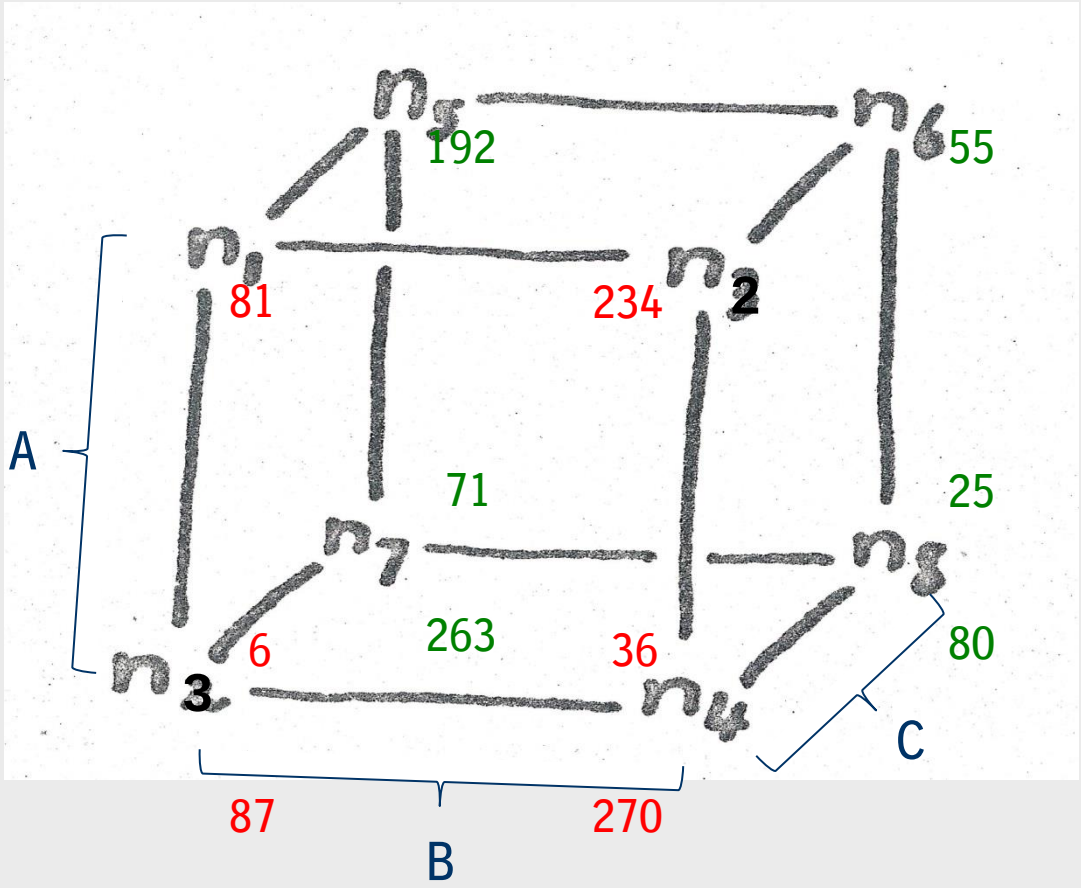
Julious/Mullee 1994: Kidney surgery.

A := success: yes/no,
B := type: open/percutaneous,
C := stone size class: small/large

(binomial model)

Boehringer Ingelheim

(b) A basic example



Julious/Mullee, B.M.J. 1994

Real examples are rare.

Yule 1903, Simpson 1951, Kendall/Stuart 1979, Chuang-Stein/ Beltangady 2011 are artificial.

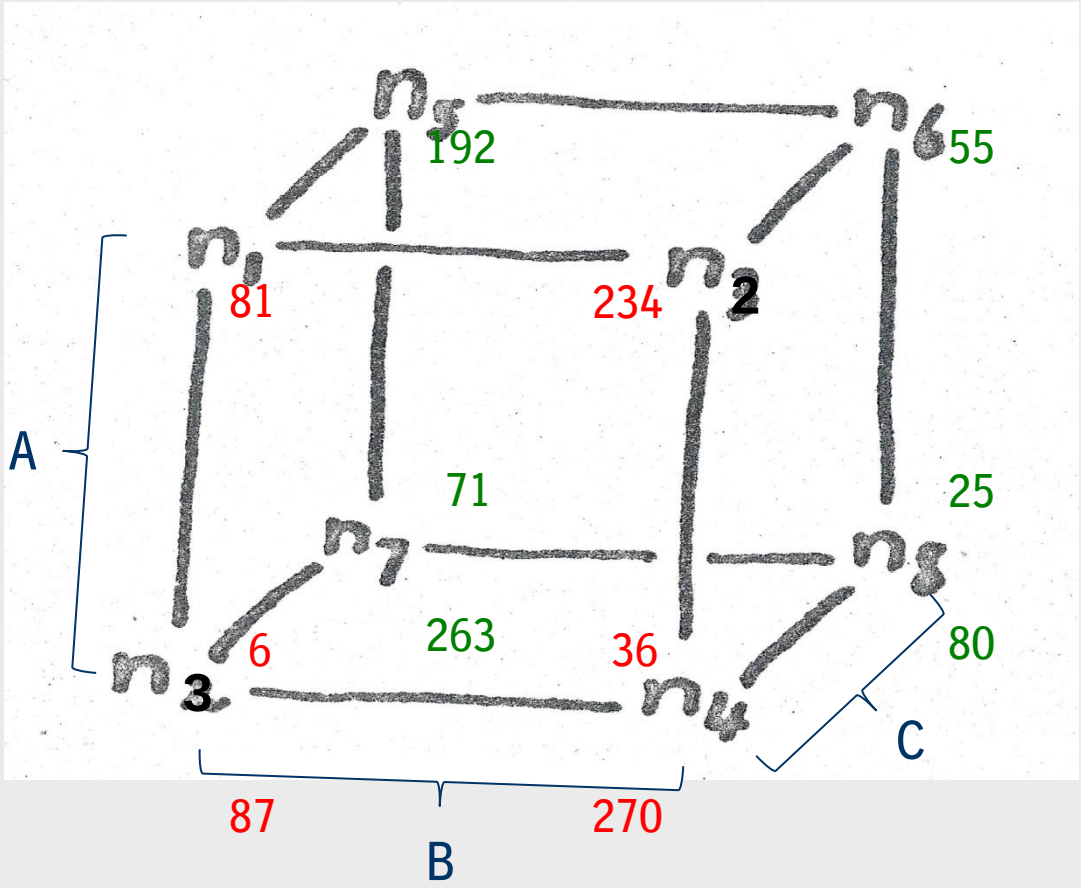
Julious/Mullee 1994: Kidney surgery.

A := success: yes/no,
B := type: open/percutaneous,
C := stone size class: small/large

(binomial model)

Boehringer Ingelheim

(b) A basic example



Julious/Mullee 1994: Kidney surgery.

A := success: yes/no,
B := type: Open/Percutaneous,
C := stone size class: small/large
(binomial model)

Est. success rates for surgery types:

O: 81/87=93.1%, 192/263=73.0%

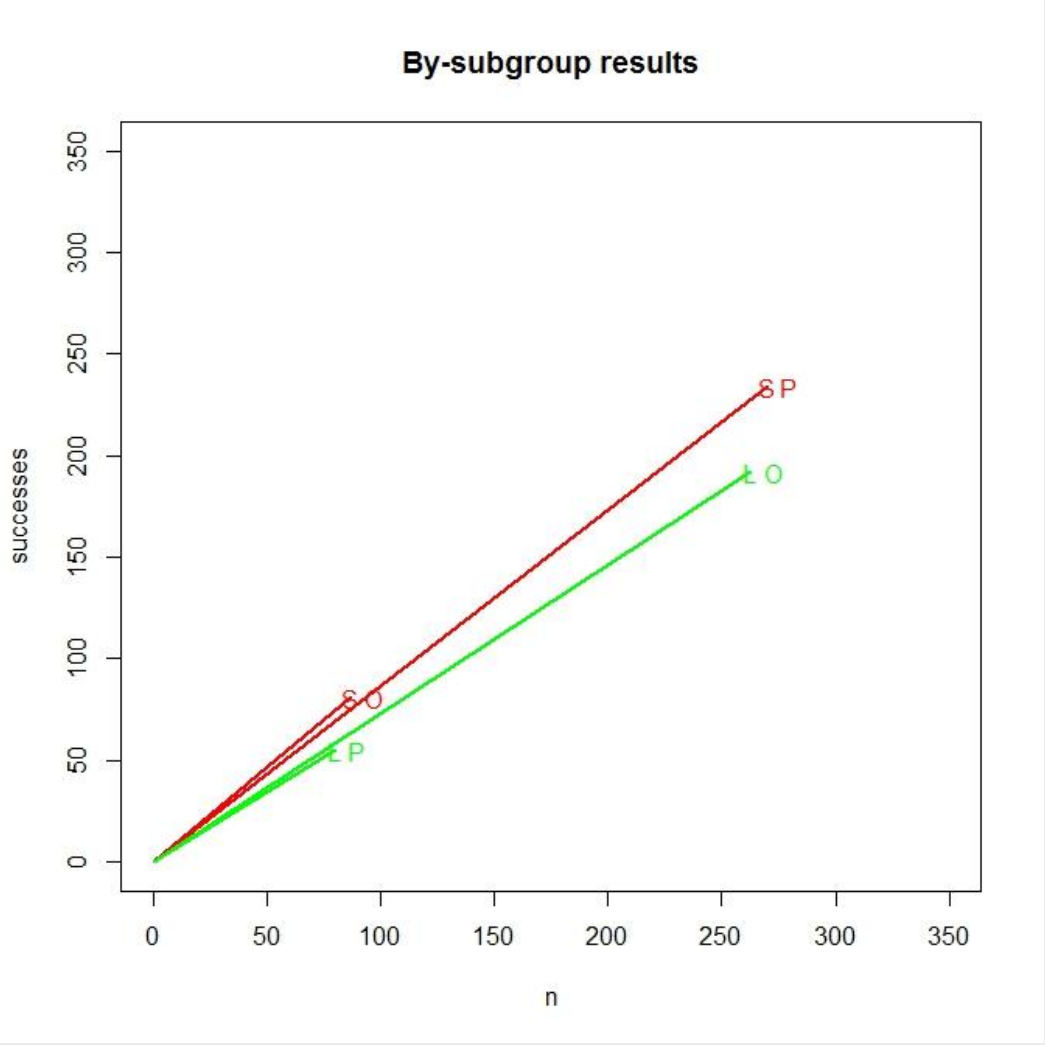
P: 234/270=86.7%, 55/80=68.8%

Together:
O: 273/350=78.0%
P: 289/350=82.6%

Julious/Mullee, B.M.J. 1994

Boehringer Ingelheim

(b) A basic example



Julious/Mullee 1994: Kidney surgery.

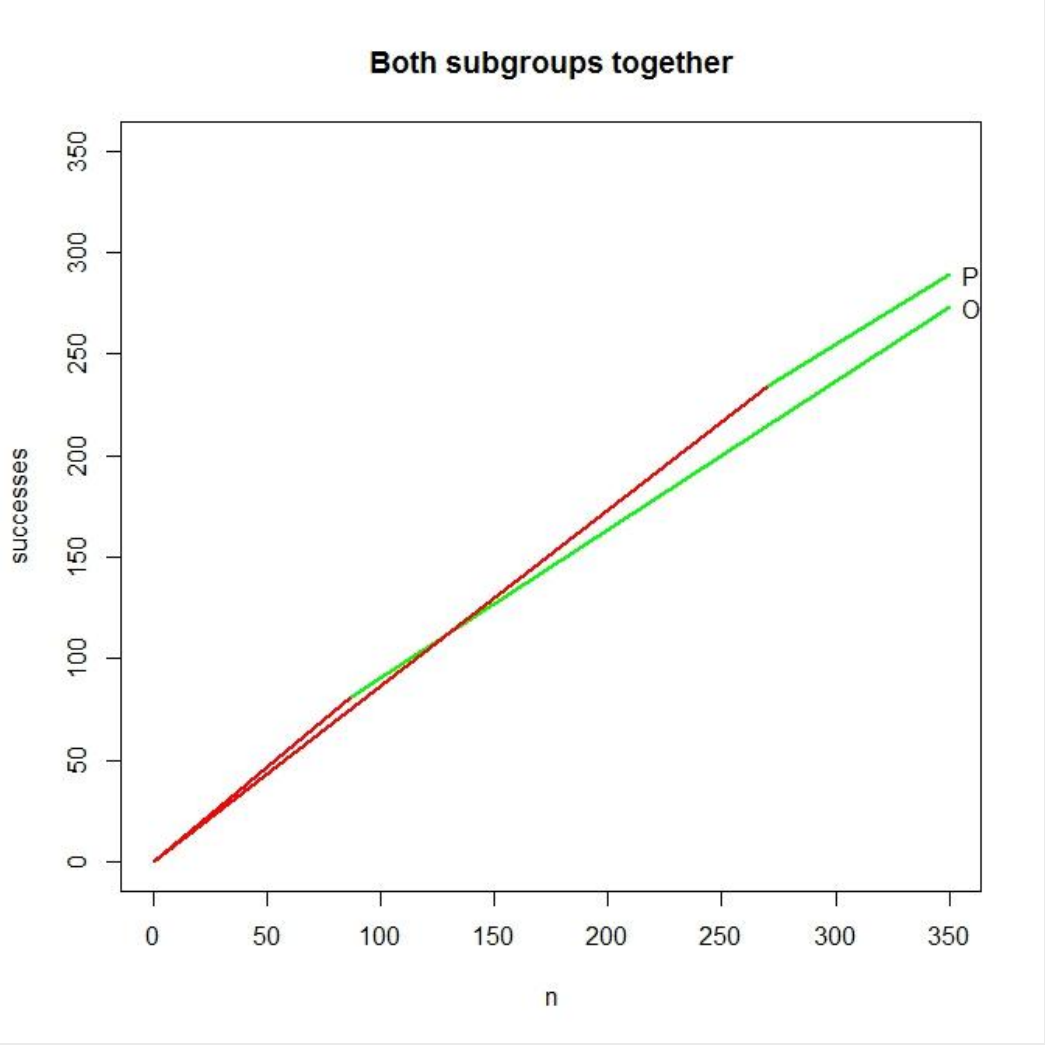
A := success: yes/no,
B := type: Open/Percutaneous,
C := stone size class: Small/Large

(binomial model)

Julious/Mullee, B.M.J. 1994

Boehringer Ingelheim

(b) A basic example



Julious/Mullee 1994: Kidney surgery.

A := success: yes/no,
B := type: Open/Percutaneous,
C := stone size class: small/large

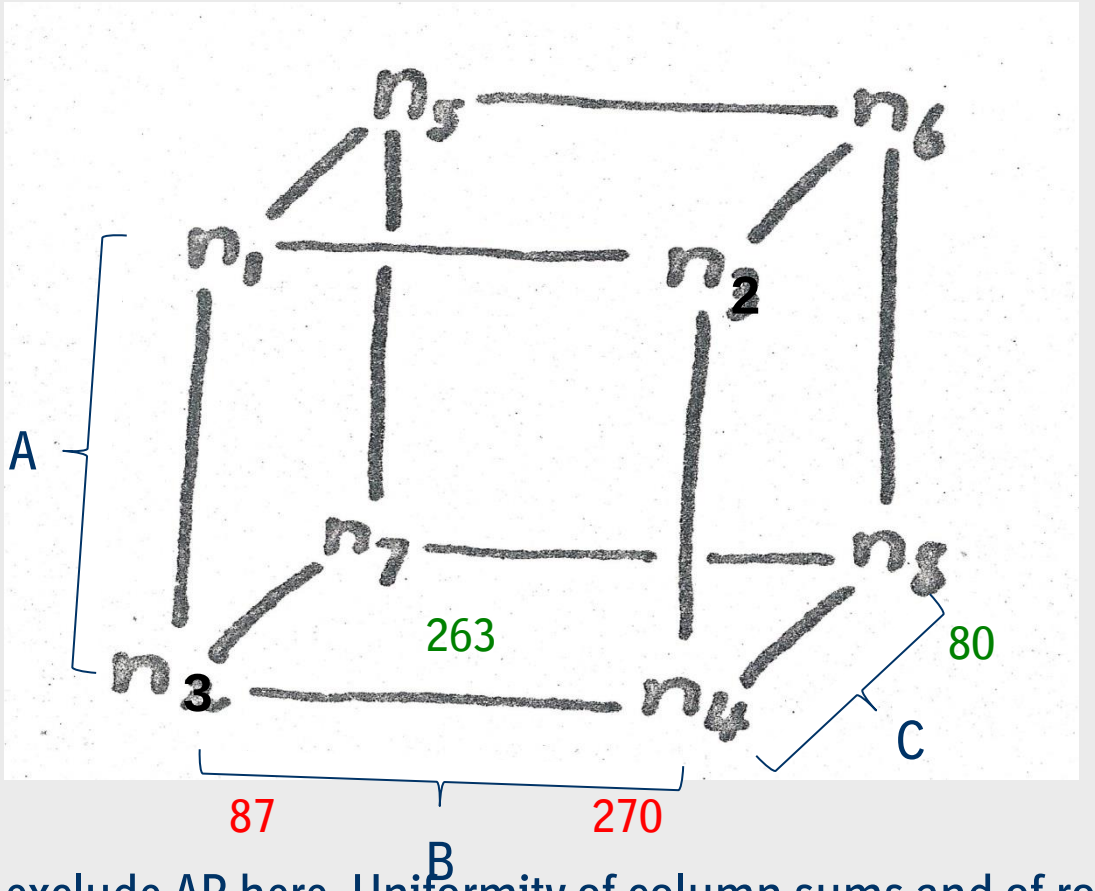
(binomial model)

After collapsing on C, we see association reversal (AR).

Julious/Mullee, B.M.J. 1994

Boehringer Ingelheim

(b) A basic example



3 classifications.

Intuitively, AR has to do with imbalance of B in the subgroups defined by C.

Good/Mittal show that if the ratio between column sums is the same for all classes of C, AR cannot occur w.r.t. the risk difference, as the marginal association will always lie in the range of the conditional associations. Corollary: Asymptotically, randomisation is sufficient to exclude AR here. Uniformity of column sums and of row sums is sufficient for absence of AR w.r.t. the OR, but none of these alone. Small deviations are permitted, and limits for these can be given.

Good/Mittal, Ann.Stat. 1987; Zidek, Biometrika 1984

(2) The prior probability for the Simpson phenomenon in the multinomial model

Boehringer Ingelheim

We go back to the multinomial model for the 2x2xK table, special case K=2, and consider an 8-tuple of probabilities $p_{1..8}$ which sum up to 1 and are naturally ≥ 0 and ≤ 1.

This 8-tuple can be interpreted as a point on the 7-dimensional „probability simplex“ in R^8.

We define the Dirichlet distribution on that simplex, with parameter tuple $\alpha_{1..8}$, as the product (up to normalization) of the $p_i^{(\alpha_i-1)}$, whereby all α_i‘s are > 0. As a special case, $\alpha_{1..8}$ = (1,...,1) gives the uniform distribution.

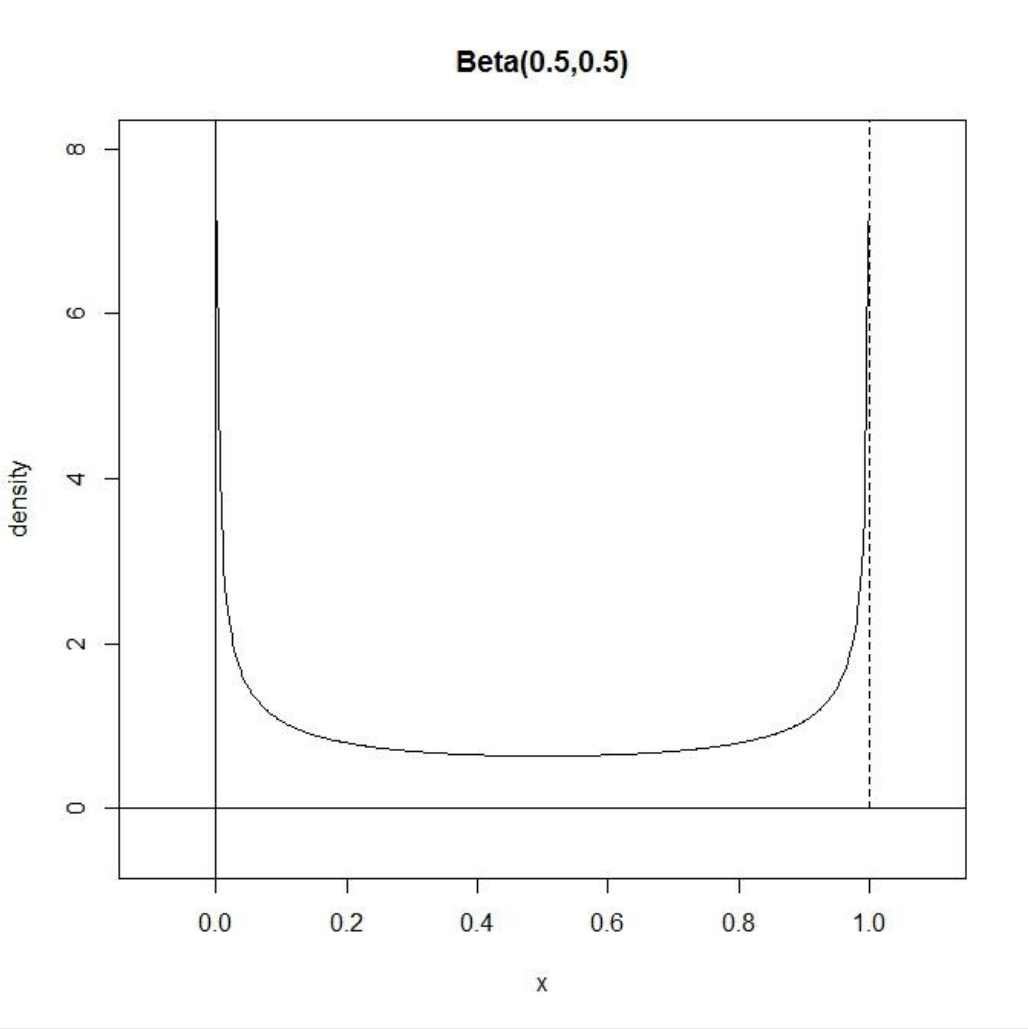
The Dirichlet distribution is conjugate to the multinomial distribution for the n_i‘s. The special case $\alpha_{1..8}$ = (0.5,...,0.5) is the Jeffreys prior distribution for the multinomial model.

Pavlides/Perlman, Am.Stat. 2009

(2) The prior probability for the Simpson phenomenon in the multinomial model

Boehringer Ingelheim

Illustration in 1 dimension:

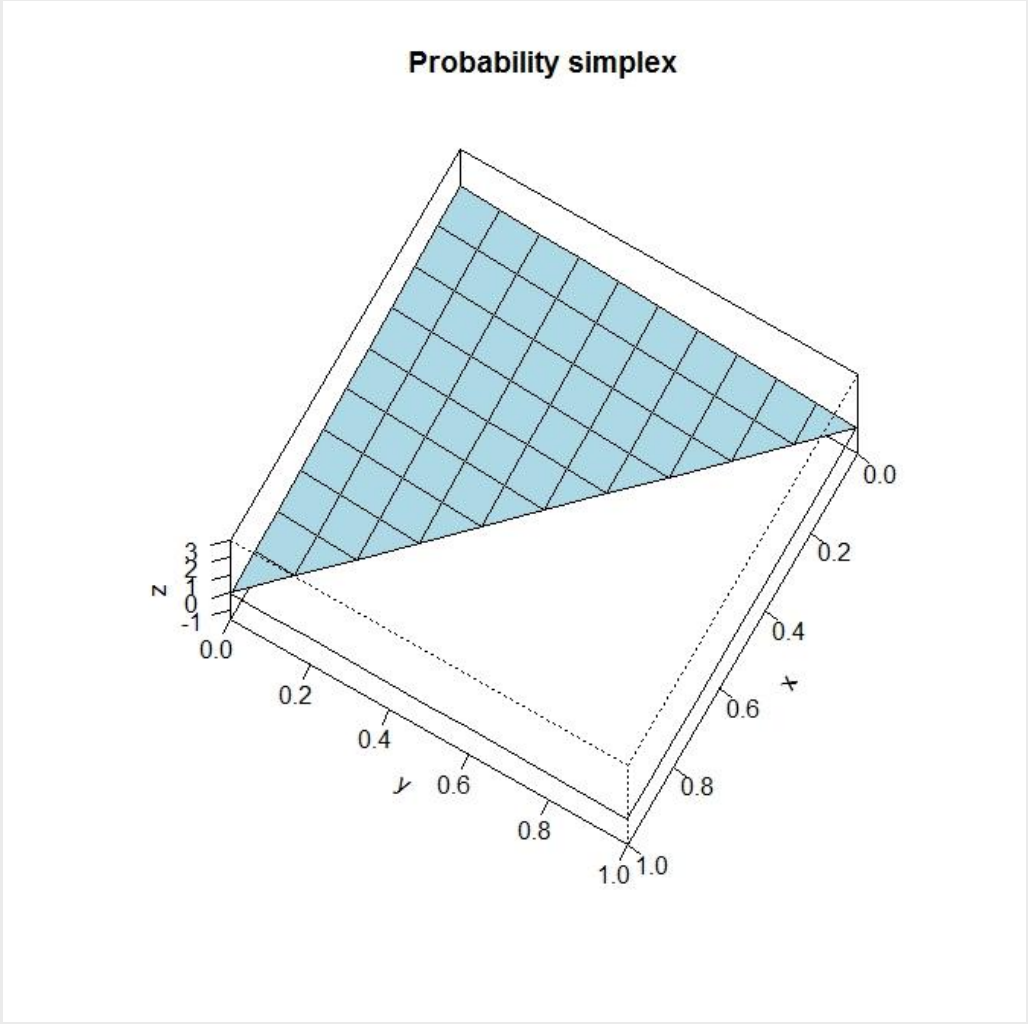


(Would have been smarter to show the 1-simplex (line from (0,1) to (1,0)) in R^2 instead of the unit interval of R^1)

(2) The prior probability for the Simpson phenomenon in the multinomial model

Boehringer Ingelheim

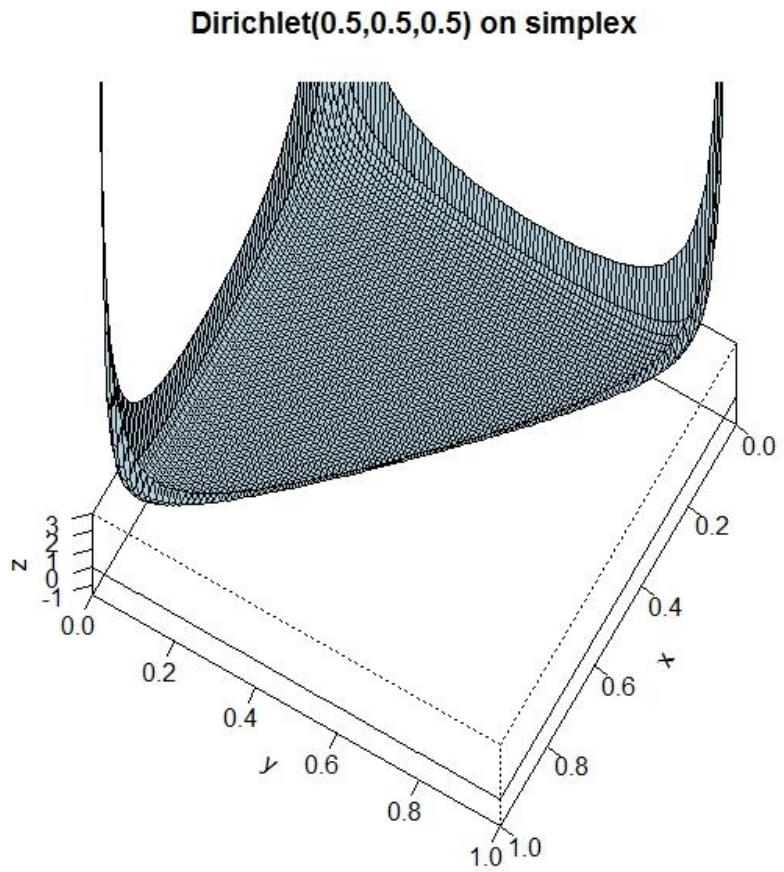
Illustration in 2 dimensions:



(2) The prior probability for the Simpson phenomenon in the multinomial model

Boehringer Ingelheim

Illustration in 2 dimensions: $\alpha_{1..3} = 0.5$

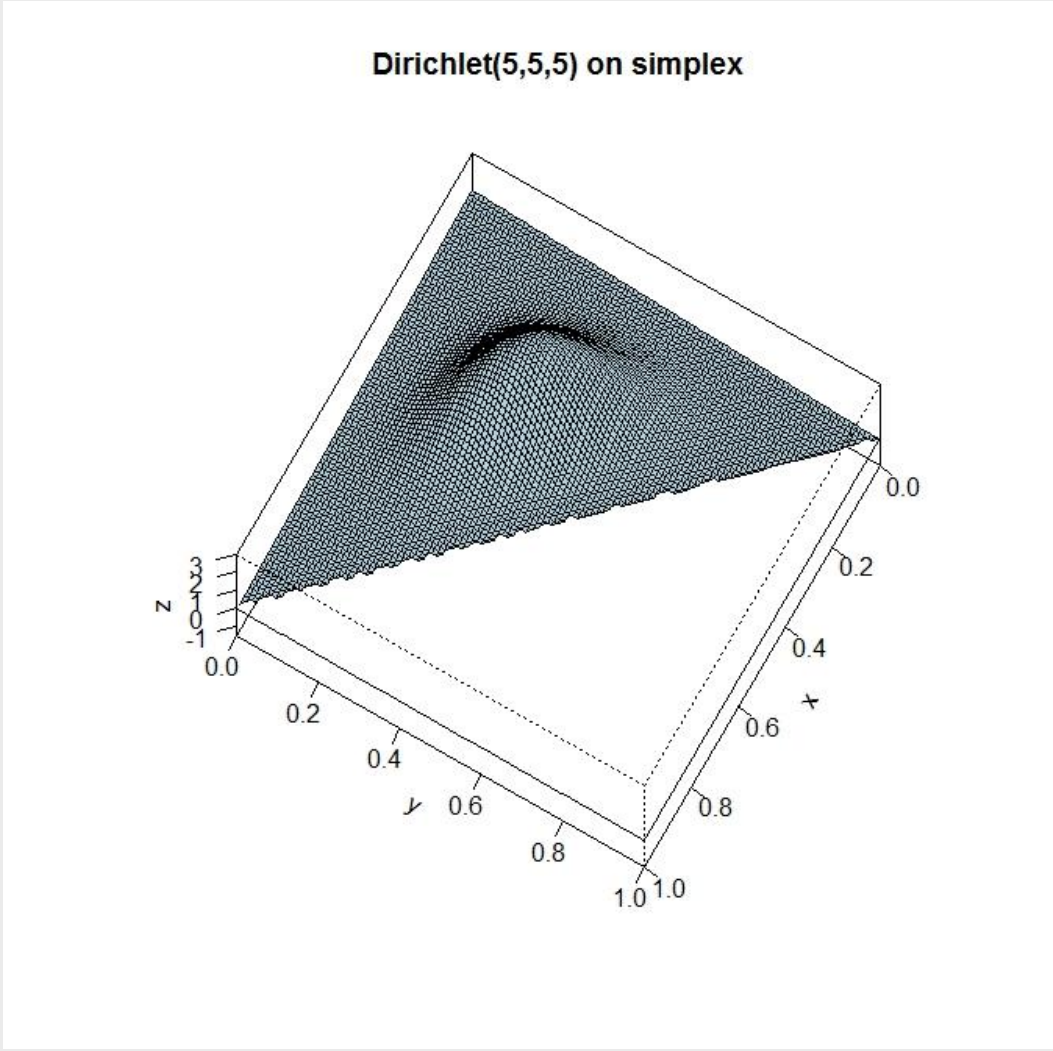


Tuples close to the boundary have a higher probability than tuples in the middle of the simplex, if $\alpha_{1..3}<1$

(2) The prior probability for the Simpson phenomenon in the multinomial model

Boehringer Ingelheim

Illustration in 2 dimensions: $\alpha_{1..3} = 5$



(2) The prior probability for the Simpson phenomenon in the multinomial model

Boehringer Ingelheim

We consider the following subset of the 7-simplex:

$$p_1 * p_4 \geq p_2 * p_3$$

$$p_5 * p_8 \geq p_6 * p_7$$

$$(p_1+p_5) * (p_4+p_8) \leq (p_2+p_6) * (p_3+p_7)$$

with at least 1 inequality strict „positive association reversal“

or all 3 inequalities inverted „negative association reversal“.

We know that the subset is not empty.

Pavlides/Perlman, Am.Stat. 2009

(2) The prior probability for the Simpson phenomenon in the multinomial model

Boehringer Ingelheim

We consider the following subset of the 7-simplex:

$$p_1 * p_4 \geq p_2 * p_3$$

$$p_5 * p_8 \geq p_6 * p_7$$

$$(p_1+p_5) * (p_4+p_8) \leq (p_2+p_6) * (p_3+p_7)$$

with at least 1 inequality strict

or all 3 inequalities inverted.

We know that the subset is not empty. Its content, weighted by a Dirichlet distribution, is the prior probability for the Simpson phenomenon, $\pi_2(\alpha_{1..8})$. It consists of 2 summands for positive and negative AR, respectively: $\pi_2^+(\alpha_{1..8})$ and $\pi_2^-(\alpha_{1..8})$.

See Pavlides/Perlman for i.i.d. MC integration based on the uniform distribution = Dir(1,...,1), on the Jeffreys distribution = Dir(0.5,...,0.5), as well as on Dir(2,...,2), Dir(3,...,3), Dir(4,...,4) and Dir(5,...,5). They also show analytically that the prior probability based on the uniform distribution is exactly 1/60.

Pavlides/Perlman, Am.Stat. 2009

(2) The prior probability for the Simpson phenomenon in the multinomial model

Boehringer Ingelheim

Remark:

The 4-fold binomial model has to be traced back to the multinomial model. It is not sufficient to just investigate on a 4-cube the subset

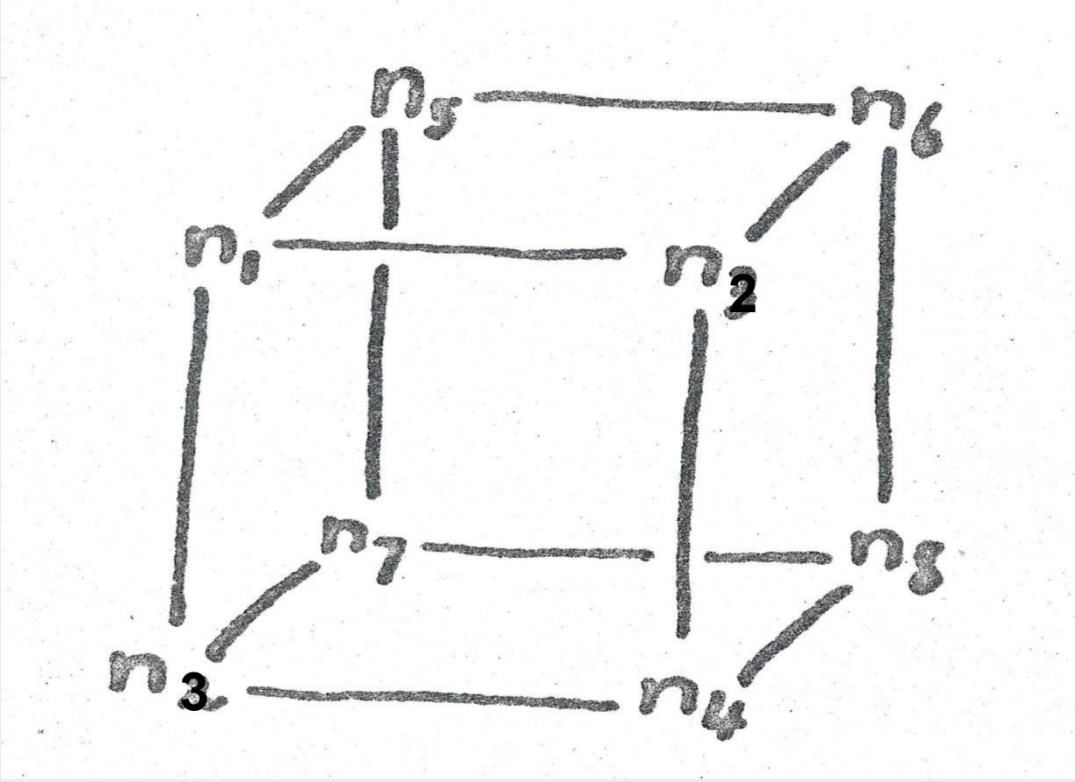
$$p_1 \geq p_2$$

$$p_5 \geq p_6$$

$$p_1+p_5 \leq p_2+p_6$$

with at least 1 inequality strict

or all 3 inequalities inverted,



as the 4 subgroup sizes – in other words, the allocation probabilities to the 4 columns – play a role as well.

Details are still open!

(3) The Bayes factor for presence or absence of the Simpson phenomenon

Boehringer Ingelheim

Let $p_{1..8}$ be a-priori distributed according to Dir(α,...,α) with $\alpha > 0$. We observe $n_{1..8}$ cases in the 8 cells of the 2x2x2 table, multinomially distributed.

Due to conjugacy, the posterior distribution of $p_{1..8}$ is then Dir($\alpha+n_1$,...,$\alpha+n_8$).

From this, we can calculate the posterior probability for that the 8-tuple $p_{1..8}$ has positive or negative AR in the same way as before.

The Bayes factor for presence of e.g. positive AR is:

Posterior odds / Prior odds

=

$(\pi_2^+(\alpha+n_1,...,\alpha+n_8)/(1-\pi_2^+(\alpha+n_1,...,\alpha+n_8))) / (\pi_2^+(\alpha,...,\alpha)/(1-\pi_2^+(\alpha,...,\alpha)))$

The example of Julious/Mullee shows negative AR. As it is based on the 4-fold binomial model, calculation of the Bayes factor is not directly possible this way – still open!

Pavlides/Perlman, Am.Stat. 2009

(4) Representation through a Directed Acyclic Graph (DAG)

Boehringer Ingelheim

Subject-matter question: When the conditional model and the marginal model give contrary answers about the association between A and B, which one is more credible?

Similar to missing-value scenarios, this is not decidable from the data alone, needs additional meta-information.

More specifically, we speak of the influence of B on A. The critical question is: Can C be associated with B and have an influence on A that does not come from B?

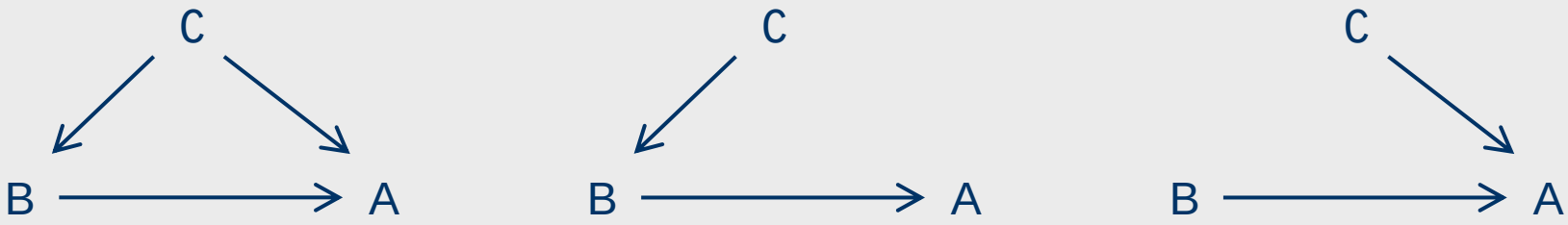
Samuels, J.A.S.A. 1993; Armistead, Am.Stat. 2014

(4) Representation through a Directed Acyclic Graph (DAG)

The directions of the influences are determined by the nature of the example.

Recap:
A = success no/yes, B = surgery type open/percutaneous, C = stone size class small/large.

Therefore, the following influences make sense empirically:



(An arrow means that influence is possible, absence means that influence is not possible)

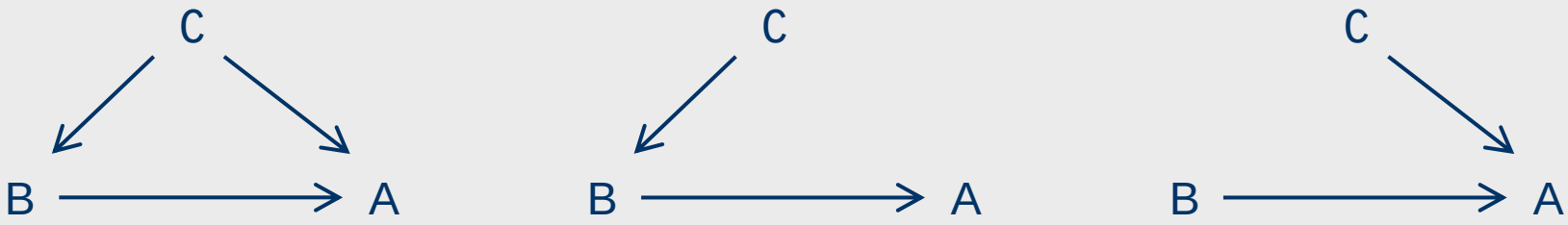
Pearl, Biometrika 1995

(4) Representation through a Directed Acyclic Graph (DAG)

The directions of the influences are determined by the nature of the example.

Recap:
A = success no/yes, B = surgery type open/percutaneous, C = stone size class small/large.

Therefore, the following influences make sense empirically:



(An arrow means that influence is possible, absence means that influence is not possible)

According to Pearl‘s „back-door“ criterion, C has to be conditioned on

In these 2 cases, C has to be ignored for the investigation of B -> A

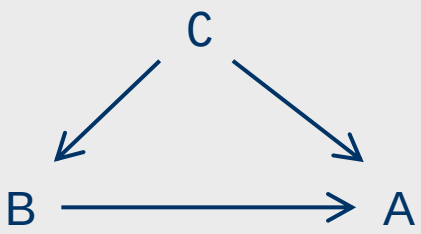
Pearl, Stat.Surv. 2009, p.114

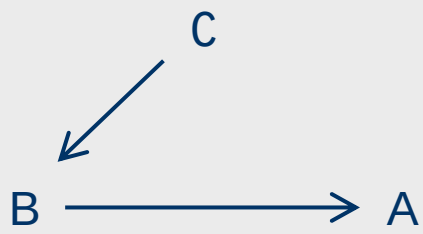
(4) Representation through a Directed Acyclic Graph (DAG)

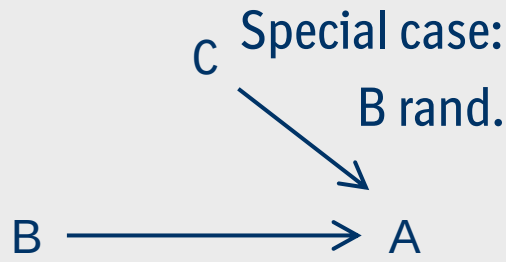
The directions of the influences are determined by the nature of the example.

Recap:
A = success no/yes, B = surgery type open/percutaneous, C = stone size class small/large.

Therefore, the following influences make sense empirically:



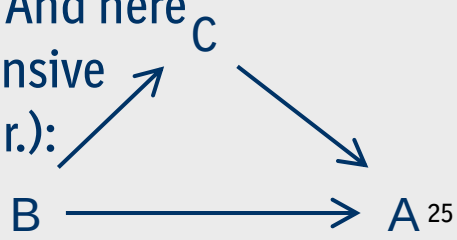




(An arrow means that influence is possible, absence means that influence is not possible)

According to Pearl‘s „back-door“ criterion, C has to be conditioned on

In these 2 cases, C has to be ignored for the investigation of B -> A. And here as well (e.g. antihypotensive trt., C := on-trt. blood pr.):



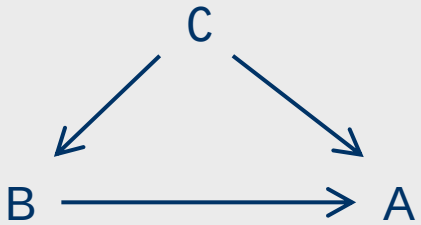
Pearl, Stat.Surv. 2009, p.114; Armistead, Am.Stat. 2014, p.5

(5) The meta-analysis example

Boehringer Ingelheim

Rücker/Schumacher re-investigate the Rosiglitazone data and show that simple addition of by-trial frequencies of Myocardial infarction leads to AR.

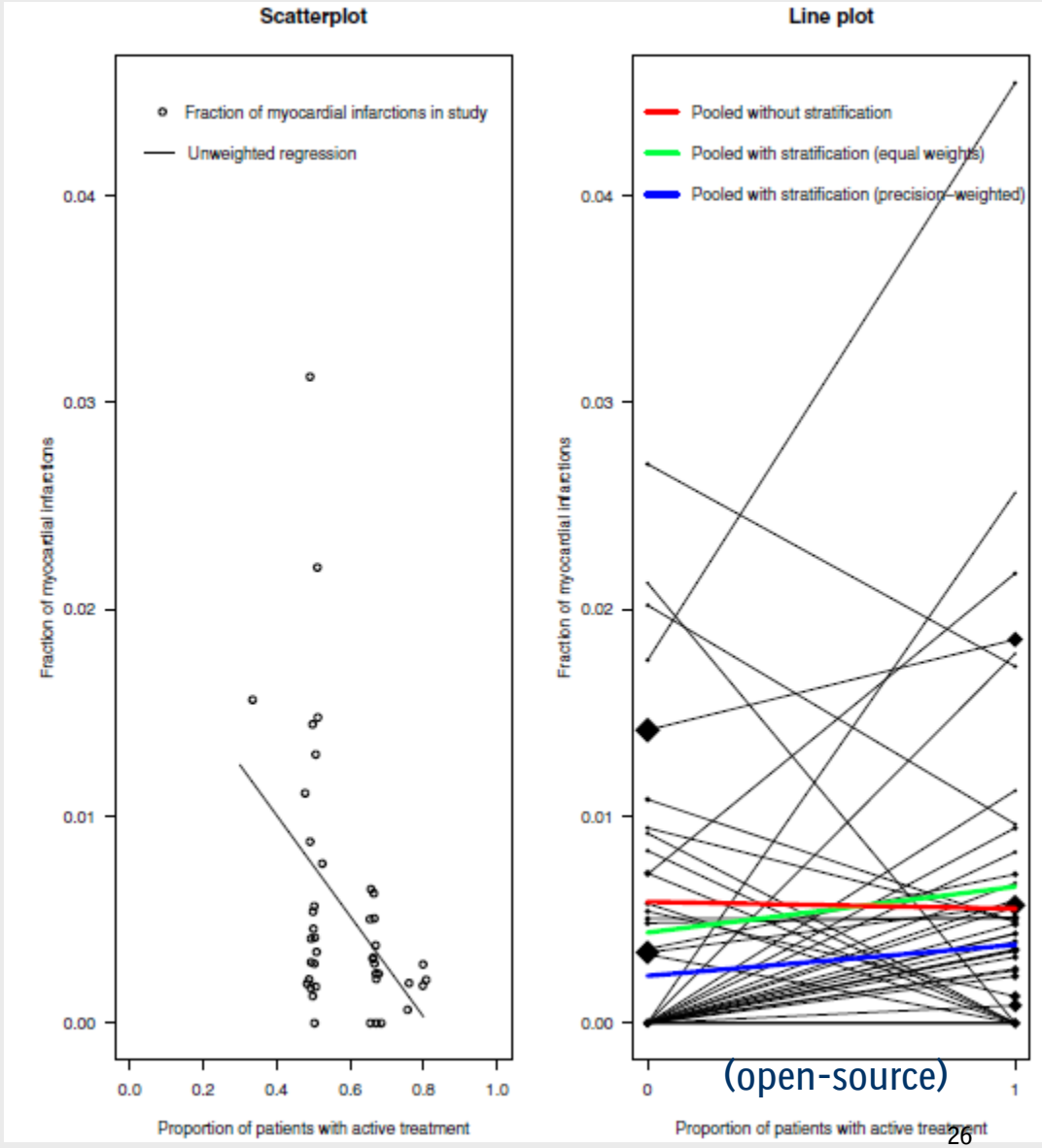
However, the influence diagram with B := treatment, C := trial:



shows that C must not be neglected and only a meta-analysis is adequate.

The same is valid for the artificial examples of Chuang-Stein/Beltangady.

Nissen/Wolski, N.E.J.M. 2007;
Rücker/Schumacher, BMC Med.Res.Meth. 2008;
Chuang-Stein/Beltangady, Pharm.Stat. 2011

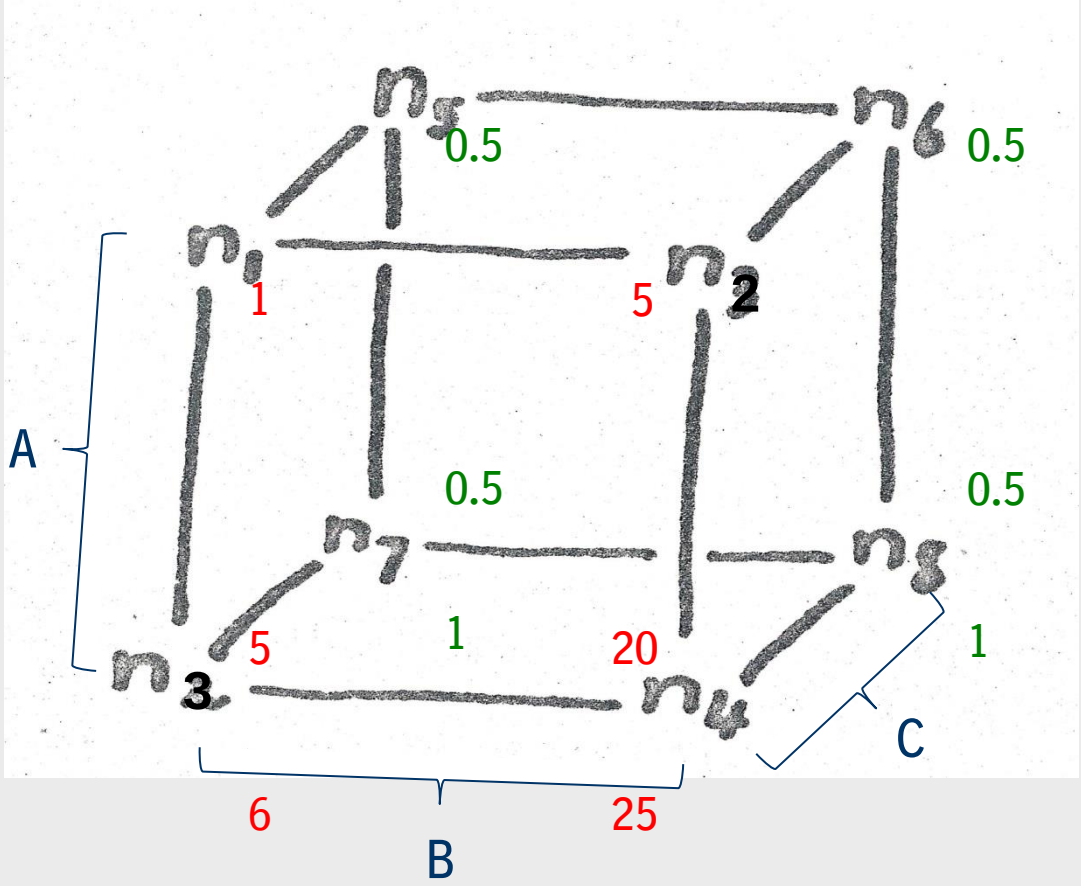


(open-source)

(6) The continuity-correction example

Boehringer Ingelheim

Greenland 2010 adds a layer of constant numbers to the 2x2 table of observed frequencies:

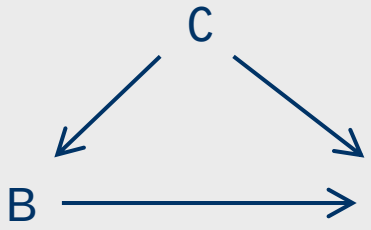


Data are artificial.

Inclusion of very small numbers makes sense as these are the situations where „continuity correction“ is actually done.

$\hat{OR}$ = 0.8, $\hat{OR}$ = 1, together 1.02.

Again, the influence of C on A and B makes the problem:

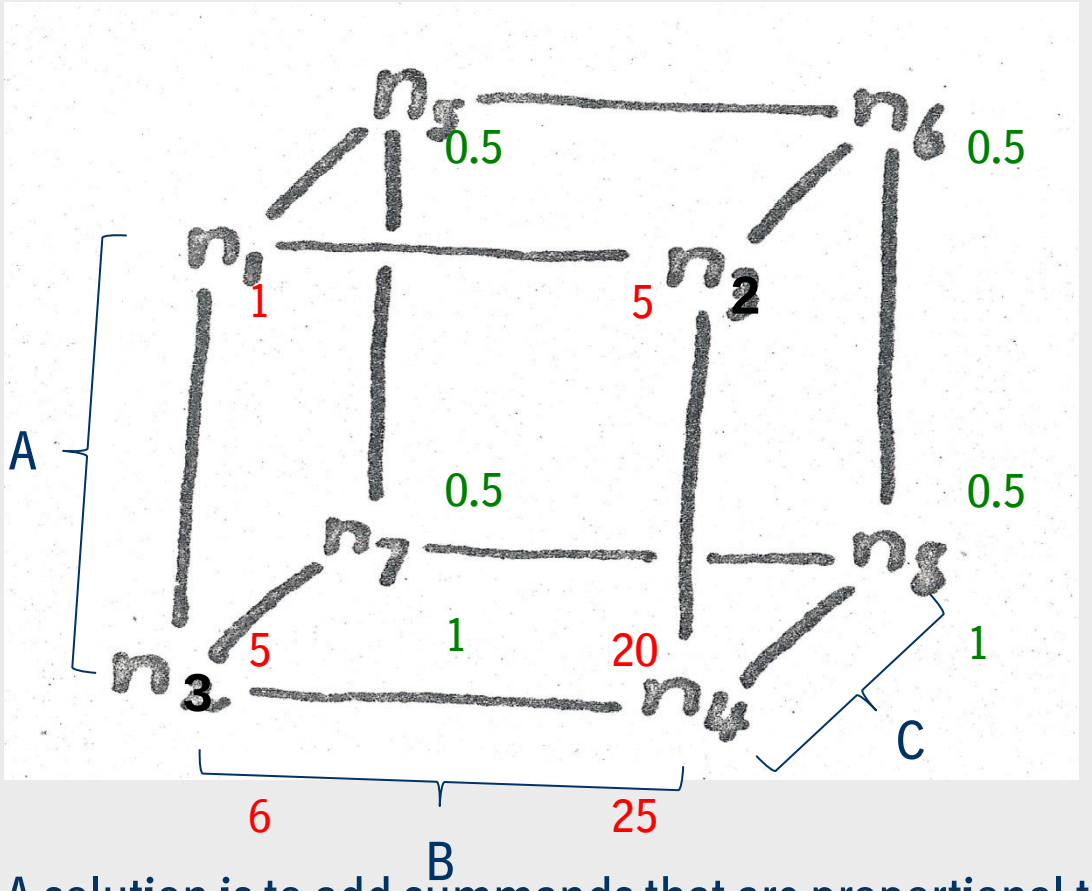


Greenland, Am.Stat. 2010

(6) The continuity-correction example

Boehringer Ingelheim

Greenland 2010 adds a layer of constant numbers to the 2x2 table of observed frequencies:

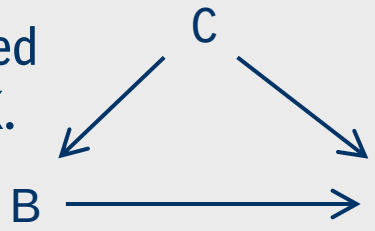


Data are artificial.

Inclusion of very small numbers makes sense as these are the situations where „continuity correction“ is actually done.

$\hat{OR}$ = 0.8, $\hat{OR}$ = 1, together 1.02.

Again, the influence of C on A and B makes the problem:



A solution is to add summands that are proportional to the expected values of the observed 2x2 table. Then shrinkage will always be OK.

Greenland, Am.Stat. 2010

(7) Discussion, outlook

Boehringer Ingelheim

The Simpson paradox can be avoided by randomisation that is independent of, or balanced w.r.t., the confounder

Its degree of certainty can be calculated for the multinomial model – for the binomial model, still open

Speaking with physicians, we should

- ask firmly for information about the nature of the confounder and about any causal relationship to the intervention
- give a clear message
- not retreat to phrases like „... has to be interpreted with caution“.

The blood-pressure example of Armistead – where finally C is to be ignored – is an example of „conditioning on a future variable“. It would be interesting to investigate similarities with NMAR modelling (selection model vs. pattern-mixture model).

Samuels, J.A.S.A. 1993, p.84; Andersen/Keiding, Stat.Med. 2012, p.1086

(8) Literature

Yule GU:
Notes on the theory of association of attributes in statistics.
Biometrika 1903; 2: 121-134

Bartlett MS:
Contingency Table Interactions.
J.R.S.S. Suppl. 1935; 2: 248-252

Simpson EH:
The Interpretation of Interaction in Contingency Tables.
J.R.S.S.B 1951; 13: 238-241

Kendall M, Stuart A:
“The advanced theory of statistics”, Vol. 2.
Charles Griffin & Co. Ltd., London / High Wycombe, 4th ed. 1979.
P. 566-575

(8) Literature

Zidek J:
Maximal Simpson-disaggregations of 2x2 tables.
Biometrika 1984; 71: 187-190

Good IJ, Mittal Y:
The amalgamation and geometry of two-by-two contingency tables.
Annals of Statistics 1987; 15: 694-711

Samuels ML:
Simpson‘s Paradox and Related Phenomena.
J.A.S.A. 1993; 88: 81-88

Julious SA, Mullee MA:
Confounding and Simpson‘s paradox.
British Medical Journal 1994; 309: 1480-1481

(8) Literature

Pearl J:
Causal diagrams for empirical research.
(With comments by Cox DR/Wermuth N, Dawid AP, Fienberg SE/Glymour C/Spirtes P, Freedman D, Imbens GW/Rubin DB, Robins JM, Rosenbaum PR, Shafer G, Sobel ME und concluding remarks by Pearl J.)
Biometrika 1995; 82: 669-710

Nissen SE, Wolski K:
Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Diseases.
N.E.J.M. 2007; 356: 2457-2471

Rücker G, Schumacher M:
Simpson‘s paradox visualized: The example of the Rosiglitazone meta-analysis.
BMC Medical Research Methodology 2008; 8(34): 1-8

(8) Literature

Pavlides MG, Perlman MD:
How Likely Is Simpson‘s Paradox?
American Statistician 2009; 63: 226-233

Pearl J:
Causal inference in statistics: An overview.
Statistics Surveys 2009; 3: 96-146

Greenland S:
Simpson‘s Paradox From Adding Constants in Contingency Tables as an Example of Bayesian Noncollapsibility.
American Statistician 2010; 64: 340-344

Chuang-Stein C, Beltangady M:
Reporting cumulative proportion of subjects with an adverse event based on data from multiple studies.
Pharmaceutical Statistics 2011; 10: 3-7

(8) Literature

Andersen,PK, Keiding N:
Interpretability and importance of functionals in competing risks and multistate models.
Statistics in Medicine 2012; 31: 1074-1088

Armistead TW:
Resurrecting the Third Variable: A Critique of Pear‘s Causal Analysis of Simpson‘s Paradox.
(With comments by Pearl J, Christensen R, Liu K/Meng X-L and concluding remarks by Armistead T.)
American Statistician 2014; 68: 1-31.
Correction: American Statistician 2014; 68: 132