University of Applied Sciences Wedel

Course of Studies: Business Informatics
Winter Term 2017/2018

Seminar: Simpson's Paradox

Paper supervised by: Prof. Dr. Iwanowski

Regarding the topic: Mathematical issues

Deadline: 19.02.2018

Author: Polonskiy, Michael

FACHSEMESTER: 5
VERWALTUNGSSEMESTER: 5
MATRICULATION NUMBER: WINF101953
E-MAIL: WINF101953@FH-WEDEL.DE

I Table of Contents

II List of Illustrations

III List of Formulae

1. Introduction

Simpson's Paradox seems to hunt statisticians for more than half a century. To this very day it is an extraordinary example for how erroneous conclusions drawn from a statistical study can be if not done correctly. It demonstrates how important it is to evaluate data carefully, with the appropriate knowledge and quality of education.

Even today some statisticians think that this paradox is unresolved, for example the authors of the article "Das Beunruhigende Paradox von Simpson" published in the German scientific journal "Spektrum der Wissenschaft"[1]. Although this journal is considered to be highly reliable the authors claim that so far no appropriate solution has been found on how to resolve the occurrences of this paradox. However, the following chapters of this paper will show that this is not true.

1.1 History of Simpson's

Simpson's paradox has been mentioned as early as 1899 by Karl Pearson, a British statistician who found a correlation between length and breadth of the human skull[2]. However, when he separated the data according to gender the correlation vanished. In 1903, statistician George Udny Yule mentioned similar effects when studying and discussing correlation and association[3]. In 1951 Edward H. Simpson was the first to address this matter in a technical paper[4]. Several years later mathematician Colin R. Blyth gave the paradox Simpson's name. Today, there exist multiple names for it, e.g. the Yule-Simpson-Effect or the Reversal-Paradox.

1.2 What is Simpson's paradox?

Simpson's Paradox is a phenomenon in Statistics in which a trend appears in different groups of data but disappears or reverses when the groups are combined.

The following example illustrates the paradox by using two different sets of cards[5]:

Two separate piles are presented, one consisting of the first deck of cards the other of the second. The goal is to choose the set with the higher overall chance of getting a red card.

[1] (cf. Delahaye 2017)
[2] (cf. Salkind 2010)
[3] (For further information see the following paper: Aldrich John 1995, Correlations Genuine and Spurious in Pearson and Yule)
[4] (To read the original paper see: Simpson E.H. 1951, The Interpretation of Interaction in Contingency Tables)
[5] (corresponding video: SciShow 2017, Statistical Paradoxes with MinutePhysics – SciShow Talk Show)

The first two piles look like this:

Pile I:
0% chance of getting a red card card

Pile II:
1/4 = 25% chance of getting a red

Figure 1: Pile I first example

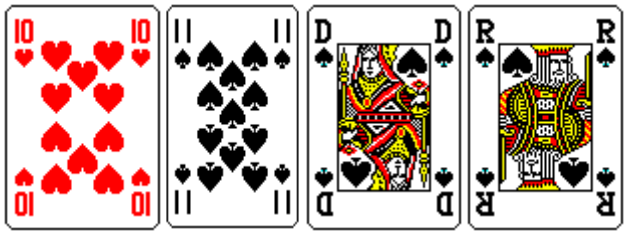


Figure 2: Pile II second example

Clearly the decision would be made in favour of the second pile since the chances of getting a red card in pile I are 0%.

Now consider the second example:

Pile I:
3/4 = 75% chance of getting a red card card

Pile II:
1/1 = 100% chance of getting a red

Figure 3: Pile I second example

Figure 4: Pile II second example

Again, the choice would be made in favour of pile II since the chance of getting a red card is 100%.

Now consider the combined piles:

Pile I (combined):
3/5 = 60% chance of getting a red card

Figure 5: pile I combined

Pile II (combined): chance of getting a red card
2/5 = 40%

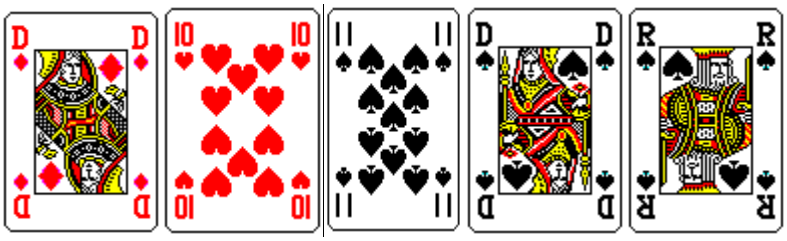


Figure 6: pile II combined

Taking into account that pile II has been both times the better choice, it should seem only reasonable that pile II will yield the better results. However, as can be seen from the graphic when the two piles are combined pile I has a better chance of getting a red card than pile II.

2 Famous Examples of Simpson's paradox:

Throughout the last decades, Simpson's paradox has caused quite a couple of headlines. This chapter will present a few of the most famous occasions of Simpson's paradox.

2.1 Berkely University Admissions[6]

In 1973 the University of California Berkely was sued for sex discrimination. The reason for this law suit was a study that showed that the average acceptance rate of females into the University was about 9% lower (35%) than the one of males (44%).

During the case, each department was investigated separately to determine the magnitude of discrimination. Surprisingly, the investigation established that each department isolated showed no signs of discrimination against females but moreover a slight tendency against males. It turned out that most women tended to apply to departments where the overall acceptance rate was very low due to a lot of applicants (departments C to F see figure 7). On the other hand, men tended to apply to departments with a relatively high overall acceptance rate (Like departments A and B). Considering this, it only appeared like women were discriminated against. Regardless of gender the general chance of getting into a department, that is more favoured by students, is lower than the chance of getting into a less popular department.

[6] (Cf. Bickel P.J. 1975, Sex Bias in Graduate Admissions)

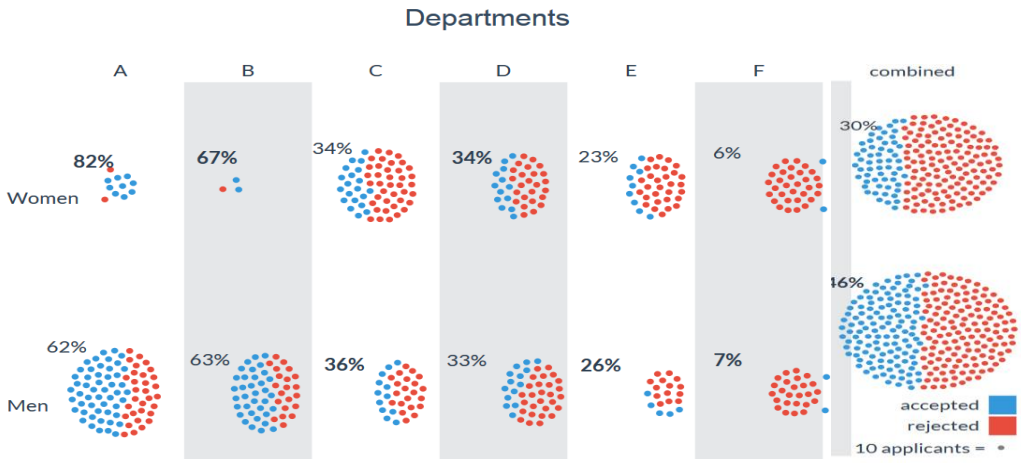


Figure 7: chart illustrating the distribution of student applications for the six major departments at Berkely

2.2 Mortality Rates for Smokers VS Non-Smokers[7]

In the early 1970s a study showed that the average mortality rate of a non-smoker (31.4%) is 7.5% higher than that of a smoker (23.9%). This suggested that smoking benefits life expectancy.

Splitting the data per age groups showed a reverse trend. In each age group the mortality rate of non-smokers was significantly less than that of smokers.

Figure 8: Aggregated mortality rates

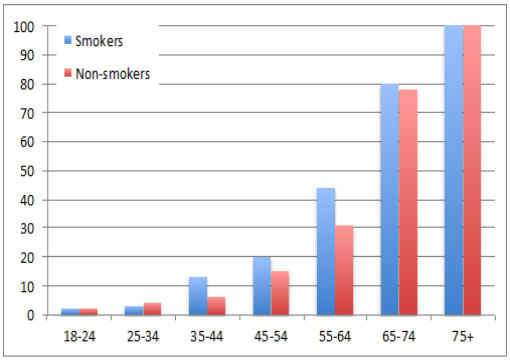


Figure 9: Mortality rates split per age

The reason for the seemingly higher mortality rate of non-smokers was the different age distribution between the two groups. Non-smokers tend to live longer than smokers, therefore the percentage of older people was significantly higher in the non-smoking group. Because older people die more often it appeared as if non-smokers had a higher mortality rate than smokers.

[7] (Cf. Schmarzo Bill 2014, Beware Simpson's Paradox)

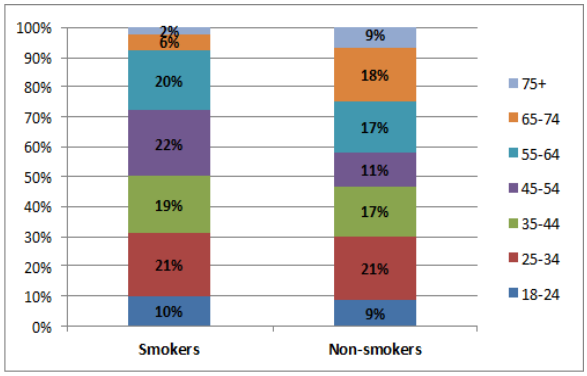


Figure 10: Age distribution of smokers and non-smokers

2.3 Kidney Stone Treatment[8]

In 1986 a medical study suggested that of two different kidney stone treatments, for simplification called A and B, treatment B (83%) was to be preferred to use. It had a 5% higher overall success rate than treatment A (78%).
However, when the data was separated according to the size of the kidney stones it showed treatment A was more effective in both groups.
This was since treatment B had been usually used on small kidney stones where the overall success rate is higher than when treating large kidney stones.

	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Figure 11: Distribution of treatment A and B regarding the size of the kidney stones.

[8] (Cf. Julious 1994, Confounding and Simpson's Paradox; Wikipedia, Simpson's paradox)

3. Graphical Visualization of Simpson's Paradox[9]

To get a better understanding of the effect it is helpful to look at the paradox depicted with vectors since in this case it is more tangible and therefore less paradoxical. The chart below shows six different vectors divided into three subgroups each paired up according to their number, blue vs red. The dimension in which these vectors are compared is their slope.

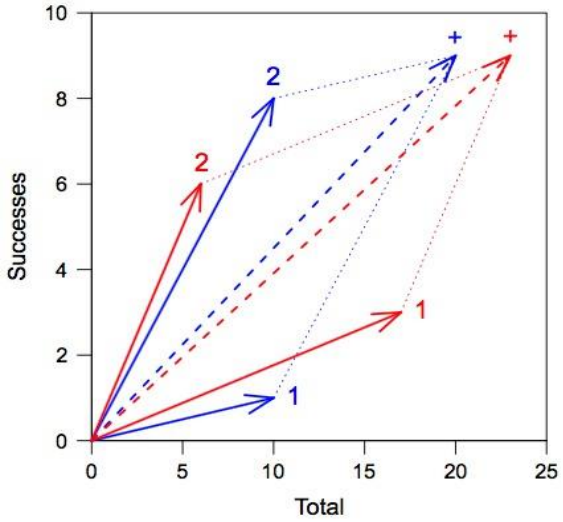


Figure 12: Six vectors, three red and three blue

red number one + red number two = dashed red with sign +
blue number one + red number two = dashed blue with sign +

Considering the two vectors at the bottom (titled with the number one) the red vector has a higher slope. Same holds for the two vectors at the top (titled with number two), the red vector has the higher slope.

However, when the two blue and the two red vectors are combined yielding the dashed red and the dashed blue vector the resultant blue vector has a greater slope than the resultant red one. This illustrates how a result can dominate in two separate subgroups (red vectors one & two > blue vectors one & two) but be reversed when the two are combined (blue dashed vector > red dashed vector).

4. How to resolve Simpson's paradox?

Considering the presented examples, the answer as to which set of data holds the correct interpretation has always been provided and explained by a logical explanation. Unfortunately, not every case of this paradox can be solved and understood that easily. Two main questions appear when dealing with Simpson's:

[9] (Cf. Wikipedia, Simpson's paradox)

1. **How to make the correct decision?**
   If confronted with two different conclusions depending on which set of data is considered (aggregated or disaggregated) which one should be chosen and what is the justification for that?

2. **How do we know if Simpsons Paradox has occurred?**
   The presence of Simpson's paradox is not always evident since it depends on the criteria which we condition upon if a case of the paradox is discovered. If for example the study regarding the gender bias at Berkely had never chosen to investigate each department separately therefore splitting up the data a reverse trend would have never been detected. Choosing reasonable criteria for grouping your data is an art which is crucial to the validity of a study. However, once the set of pertinent criteria is chosen there are ways to determine if a case of Simpson's paradox appears or not.

4.1 How to make the correct decision?

When two different conclusions are provided, depending on which kind of partitions are made, the variables involved need to be modelled in a causal diagram.

To decide which set of data holds the correct answer the following approach has been developed and introduced by Judea Pearl, professor for computer science and statistics at the University of California Los Angeles. Pearl uses the concept of causality and causal diagrams to explain how to resolve a case of Simpson's Paradox[10].

4.1.1 An introduction to causality

A causal diagram consists of two things[11]: nodes and edges. Each node represents a variable that can take on a value from a finite set of values, e.g. rain: (true, false), grass (wet, dry). Every node in the diagram must be connected to at least one other node by an edge. Each edge has a distinct direction representing the causal relationship between two nodes, e.g. rain causes the grass to become wet.

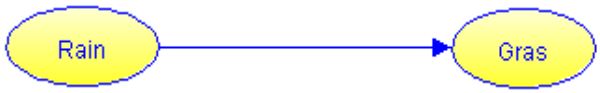


Figure 13: simple causal diagram

When saying, rain causes grass to become wet we have to define the kind of causation used in this term.

Causality differentiates between two kinds of relationships[12]:

Deterministic causation: If A causes B, than A must always be followed by B.

[10] (Cf. Pearl 1995, Causal diagrams for empirical research)
[11] (Cf. Yudkowsky 2012, Causal Diagrams and Causal Models)
[12] (Cf. Wright 2008, Types of Causality)

> For Example: Applying heat to water will cause the water to become warmer and eventually it will start boiling.

And probabilistic causation: A probabilistically causes B if A's occurrence increases the probability of B's occurrence.

> For Example: Smoking increases the risk of lung cancer. However, smoking will not cause lung cancer with a 100% certainty since there are people who smoke all their life but have no lung cancer.

In our first example the relationship between rain and grass the type of causality can safely be titled as deterministic since all other things being equal rain will always cause the grass to become wet.

However, for dealing with Simpson's paradox probabilistic causation will have to be applied since the deterministic approach is very hard to verify because a 100% certainty is required.

4.1.2 Causation and Correlation[13]

A very important aspect to keep in mind when dealing with causation is the difference between correlation and causation. Although at first glance especially probabilistic causation may seem to have no difference with correlation it is crucial to understand that the two are fundamentally different.

A correlation between variables may be based on a causal relationship but this does not have to be the case.
Correlation does not imply causation.

Imagine the following scenario[14]:

A Barometer hanging in a house close by the shore is used to indicate whether a storm is about to come. If the level of Mercury in the Barometer starts to drop rapidly, chances are a storm is about to come up soon. This is a correlational relationship since observing the Barometer tells you something about the probability of the storm.

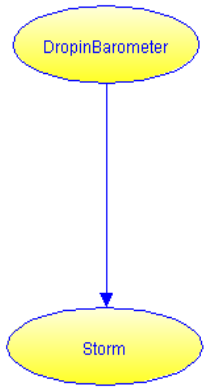
To test whether this is a causal relationship, try to actively manipulate the level of mercury in the Barometer, e.g. by applying heat or cooling, and see if the chances of a storm change accordingly. Reason tells us they will not. Hence, this is a correlation but the Barometer does not cause the storm.

By considering the air pressure, we obtain a causal relationship. If we could change the air pressure in a particular region the chances of a storm would change accordingly.

This means that air pressure influences both, storm and barometer, in a causal way whereas the relation between the barometer and the storm is of a non-causal type.

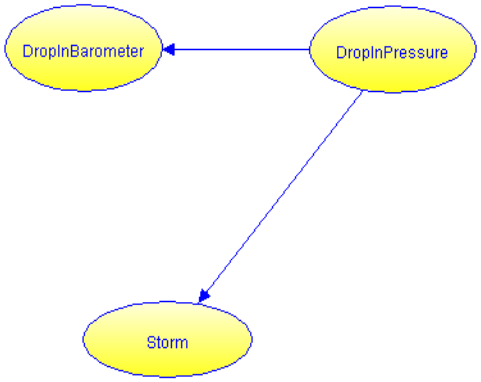
[13] (Cf. Hitchcock 2010, Probabilistic Causation)

[14] (Cf. Wikipedia, Probabilistic Causation)



Correlation

Figure 14: Example for correlation



Causation

Figure 15: Example for causation

The essential difference comes from obtaining knowledge in two different ways. In the first case knowledge is gained by observation (correlation) whereas in the second case knowledge is gained by active interference (causation).

4.1.3 Pearls Causality[15]

To express the difference between merely observing a dependence contrary to actively testing it Judea Pearl invented a new form of algebra specifically suited for the purpose of operating with causality. This algebra uses the do-operator which explicitly states that a certain probability is observed when externally intervening[16].

In these terms:

P(storm | do (air pressure falling))	describes the probability of a storm coming when the air pressure is actively lowered.

However the probability of:

P(storm | do(Barometer falling))	will yield to 0%

whereas the standard conditional probability:

P(storm | Barometer falling)	will give a value greater 0%.

Using this new calculus with its new stated laws, it is possible to model a case of Simpson's paradox and solve the equations respectively. The result will tell in which group of data the correct answer presides. For further information on this confer to Pearl's book: "Causality, Models, reasoning and Inference"[17].

This paper will not elaborate on this topic any further since using causal diagrams solves Simpson's paradox in a much easier way.

[15] (For more detailed information confer to: Pearl 2009, Causality, Models, Reasoning and Inference)
[16] (Cf. Pearl 2012, The Do-Calculus Revisited)
[17] See 15

4.1.4 Making the correct decision

Pearl uses this technique to demonstrate graphically how a case of Simpson's can be modelled and solved by applying a certain criterion to the resulting diagram[18].

The criterion used derives itself from two separate criterions which are named:
1. D-separation criterion
2. Back-door criterion

This paper will not elaborate on these two any further since there is sufficient other work that does so[18]. Instead they will be simplified in the following to the necessary criterion for determining in which set of data the correct answer lies.

When trying to identify the causal influence of a cause X on an effect Y consider the following:

Every path (sequence of nodes and edges where the direction is irrelevant!)
that contains an arrow into the cause X needs to be inspected.
If this path contains a chain: I -> m -> j
or a fork: I <- m -> j
than the node in the middle of the chain (m) or the source of the fork (m) needs to be conditioned on.

It is important to realise that a chain can occur in any direction as long as both edges go the same way, when applying this criterion.

Furthermore, if the path contains an inverted fork (called a collider): I -> m <- j
than the node in the middle (m) and all its descendant nodes are not allowed to be conditioned on.

To illustrate this, four examples are given in the following. In each example the question arises whether a decision should be based on data separated according to a criterion Z.

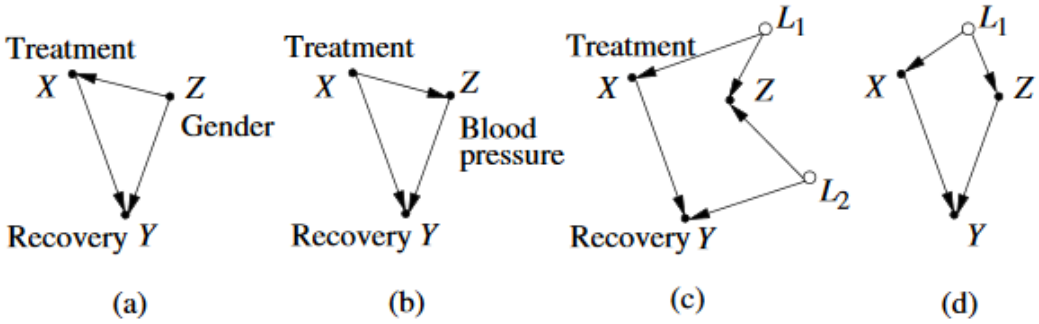


Figure 16:causal diagrams for Simpson's

Node X represents one of two treatments which have a causal effect on recovery Y when given to patients. In Example a) the doctor has two different studies, one suggesting that if he splits up his patients according to gender treatment A is to be overall preferred.

[18] (Cf. Pearl 2013, Understanding Simpson's Paradox)

The other one implying that when not considering the gender of the patient, treatment B is the better choice. The corresponding causal diagram shows that when applying the criterion, the correct answer is provided by the disaggregated data. Since X <- Z -> Y is a path with an arrow into the cause X and the path contains a fork, Z should be conditioned on.

Example b) suggests the same scenario although this time the patients are divided with regards to their blood pressure. If the blood pressure is taken into account, both groups (low and high pressure) suggest another treatment as if not considering blood pressure at all. Looking at the corresponding causal diagram shows that the correct answer is provided by the aggregated data. Thus, blood pressure should not be examined. There is no path with an arrow into the cause X, therefore no conditioning can be done.

Example c) presents multiple nodes influencing each other. Applying the criterion there is a path with an arrow into the cause (X <- L1 -> Z <- L2). However, it is forbidden to condition on Z or any of its descendants since Z is a collider. The edges from L1 and L2 collide at Z.

Example d) displays a case with a typical chain. There is a path (X <- L1 -> Z -> Y) with an arrow into the cause X that contains a chain with the node Z in the middle. Therefore, Z should be conditioned on. Another option would be to condition on L1 since it forms a fork with X and Z. This is equally possible and depends on whether all the information needed is available. For example, if the data does not provide enough information about criterion Z, same samples might lack this information, criterion L1 could be the better choice. Both ways will yield the same correct result.

Summarizing this approach, the harder task is to model a specific case precisely in a certain diagram and not to solve the paradox when given such a diagram.

4.2 How to detect a case of Simpson's Paradox

The second question is: How do we know if our data contains a case of Simpson's paradox?

When choosing the relevant criteria for a certain study it is important to be very careful and possess thorough knowledge about the task at hand. Otherwise Simpson's Paradox shows, that each result can be intentionally reversed simply by splitting the examined data according to some deliberate criterion. However, even if all criteria picked have their validation and significance, a certain sub-division can yield a case of reversal.

To be aware of such a case it is possible to apply the same approach used in the previous chapter. The theory of graphical models can tell for a specific causal diagram whether a case of Simpson's paradox is possible or not. For further information see "Graphical Models" published in 1996 by Steffen L. Lauritzen.

In the next chapter an algorithm that can be used for computing if a case of Simpson's paradox exists in the given data once the relevant criteria are established will be presented.

The algorithm has its roots in data mining since it operates on a flat data structure meaning it uses only a single table as a data source. It has been presented first by Alex A. Freitas in 1998[19].

4.2.1 Algorithm for detecting Simpson's paradox in data mining[20]

Input: The algorithms input consists of a list LG of binary user-defined goal attributes and a set of obtained data presented in a flat data base (single table).

Output: The algorithms output consists of all instances of Simpson's paradox found meaning it will provide a table of three columns:
relevant goal attribute G, first partition attribute A1, second partition attribute A2
filled with the corresponding data that lead to the occurrence of Simpson's paradox.

Restrictions: Every attribute in LG has to be binary.
Every attribute in L1 has to be binary.
Every attribute in L2 has to be categorical.
No attribute in LG can be put in L1 or L2.

Variables: LG: list of user defined binary goal attributes
L1: list of 1stPartAtt
L2: list of 2ndPartAtt
G1, G2: specific goal attribute representing the current state in population 1 and 2
Gij: specific goal attribute representing current state in a subpopulation
i=1stPartAtt
j= 2ndPartAtt
Pr(Gij): probability of the occurrence of a certain goal attribute in a subpopulation devided according to 1stPartAtt i and 2ndPartAtt j.
Pr(G1= 'yes' | A1=1): probability of the occurrence of a certain goal attribute in population one under the condition that the current 1stPartAtt A1 is equal to one.

The following part elaborates on the algorithm by Freitas and provide an example to make its functionality clearer:

1) INPUT: list of binary user-defined goal attributes L_G
2) BEGIN
3) identify attributes that can be used as 1stPartAtt and put them in list L_1
4) identify attributes that can be used as 2ndPartAtt and put them in list L_2
5) FOR EACH goal attribute G in L_G
6) FOR EACH attribute A_1 in L_1
7) partition population into Pop_1 and Pop_2, according to values of A_1

[19] (Cf. Freitas 1998, On objective measures of rule surprisingness)
[20] (Cf. Freitas 2000, Discovering Surprising Patterns by Detecting Occurences of Simpson's Paradox)

8) $Pr(G_1) = Pr(G=„yes“|A_1=1)$
9) $Pr(G_2) = Pr(G=„yes“|A_1=2)$
10) FOR EACH attribute A_2 in L_2 such that $A_2 \neq A_1$
11) FOR i=1,2
12) partition Pop_i into m new populations $Pop_{i1} \ldots Pop_{im}$,
13) according to the values of A2
14) FOR j=1,…,m
15) $Pr(G_{ij}) = Pr(G=„yes“|A_1=i,A_2=j)$
16) IF ($Pr(G_1) > Pr(G_2)$ AND $Pr(G_{1j}) \leq Pr(G_{2j})$, j=1,…,m)
17) OR ($Pr(G_1) < Pr(G_2)$ AND $Pr(G_{1j}) \geq Pr(G_{2j})$, j=1,…,m)
18) report the occurence of the paradox to the user

Imagine the following scenario[21]:

A patient comes to a doctor and has a certain disease. The doctor is informed about a study regarding this disease that yields the following table:

	Cured	Not cured	Succes rate
Medicine A	5	6	45.45%
Medicine B	4	5	44.44%

Figure 17: Combined table doctor patient example

According to this study the patient should be provided with the Medicine A.
The corresponding data for this study looks like this:

No.	Med. type	Cured	Gender	No.	Med. type	Cured	Gender
1	Med. A	Yes	M	11	Med. A	No	F
2	Med. A	Yes	F	12	Med. B.	Yes	M
3	Med. A	Yes	F	13	Med. B.	Yes	M
4	Med. A	Yes	F	14	Med. B.	Yes	F
5	Med. A	Yes	F	15	Med. B.	Yes	F
6	Med. A	No	M	16	Med. B.	No	M
7	Med. A	No	M	17	Med. B.	No	M
8	Med. A	No	M	18	Med. B.	No	M
9	Med. A	No	F	19	Med. B.	No	M
10	Med. A	No	F	20	Med. B.	No	F

Figure 18: Sample data for doctor patient example

[21] (Cf. Delahaye 2017, Das Beunruhigende Paradoxon von Simpson)

Applying the presented algorithm would look like this (the line numbers of the algorithm will be used to indicate at which step which calculations are made):

Note: For this example, we will only inspect the case L1 = Medicine Type and L2= Gender since it demonstrates how the detection works. The other case would be run by the algorithm as well

1) LG = {Cured (Yes, No)}
3) L1 = { Medicine Type (A, B) , Gender (M, F) }
4) L2 = { Gender (M, F), Medicine Type (A, B) }
5) G = Cured
6) A1 = Medicine Type
7) $Pr(Cured_A) = Pr(Cured = \text{“yes”} \mid Medicine\ Type = A)$
8) $Pr(Cured_B) = Pr(Cured = \text{“yes”} \mid Medicine\ Type = B)$
9) A2 = Gender
11) FOR I = A,B
12) $Pop_I = Pop_{IM}$ & Pop_{IF}
14) FOR J = M,F
15) $Pr(Cured_{IJ}) = Pr(Cured = \text{“yes”} \mid Medicine\ Type = I, Gender = J)$
16) IF ($Pr(Cured_A) > Pr(Cured_B)$ AND $Pr(Cured_{AJ}) \leq Pr(Cured_{BJ})$, J = M,F)
17) OR ($Pr(Cured_A) < Pr(Cured_B)$ AND $Pr(Cured_{AJ}) \geq Pr(Cured_{BJ})$, J=M,F)

Running through these steps will yield the following tables:

Gender: M	Cured	Not cured	Succes rate
Medicine A	1	3	25%
Medicine B	2	4	33%

Gender: W	Cured	Not cured	Succes rate
Medicine A	4	3	57%
Medicine B	2	1	66%

Figure 19: Split table doctor patient example

Executing the last two lines of code with this data looks like this:
<u>First iteration J = M</u>:
If $5/11 > 4/9$ AND $1/4 \leq 4/7$ -> T
OR $5/11 < 4/9$ AND $1/4 \geq 4/7$ -> F

<u>Second iteration J = F:</u>
If $5/11 > 4/9$ AND $2/6 \leq 2/3$ -> T
OR $5/11 < 4/9$ AND $2/6 \geq 2/3$ -> F

Since one of the two expressions evaluates to true, a case of Simpson's paradox is detected and will be reported to the user in the following way:

Simpson's	Goal Attribute	1stPartAtt	2ndPartAtt
Yes	Cured	Medicine	Gender

Figure 20: Sample result of detection algorithm

This informs the doctor that for the data given there is a case of Simpson's paradox to be aware of if sorting the data first regarding the type of medicine and afterwards for gender. Using the method introduced in the previous chapter the doctor has now the means to determine which set of data needs to be considered.

For everybody further interested in this method the paper published by Carem C. Fabris and Alex A. Freitas: "Discovering Surprising Instances of Simpson's Paradox in Hierarchical Multidimensional Data"[22] can be recommended. It introduces the reader to an altered version of this algorithm fit to be used in relational databases. Moreover, the authors add the possibility to rank the degree of surprisingness a single instance of Simpson's paradox provides discussed in the next chapter.

4.2.2 Magnitude of surprisingness[23]

The magnitude of surprisingness M of a single scenario can be determined by measuring the degree to which the probability of the result reverses when splitting the data according to a criterion.

Imagine the following scenario: The data in question recommends taking medicine A with a 90% recovery rate versus medicine B with a 45% recovery rate. Once the data is split regarding gender medicine B is recommended with 90% versus medicine A with 45%. This would be considered a very strong case of Simpson's paradox. But if the probabilities are A: 45% and B: 44% and once split by gender A: 43%: B: 45%, than the magnitude is not very large.

The formulas for calculating this degree are the following two:

1) $M = (M1 + M2) / 2$

Equation 1: Formula for magnitude of surprisingness

2) $M1 = \frac{|\Pr(G1)-\Pr(G2)|}{\max(\Pr(G1),\Pr(G2))}$

Equation 2: Formula for magnitude of 1stPartition

$M2 = \sum_{k=1}^{m} \left(\frac{|\Pr(G1k)-\Pr(G2k)|}{\max(PrG1k),\Pr(G2k))} \right) / m$

Equation 3: Formula for magnitude of 2ndPartition

Formula 1) states that the overall magnitude is the arithmetic average of the two sub-magnitudes of the different subsets.

[22] (Cf. Freitas 2006, Discovering Surprising Instances of Simpson's Paradox in Hierarchical Multidimensional Data)
[23] (Cf. Freitas 2000, Discovering Surprising Patterns by Detecting Occurences of Simpson's Paradox)

Formula 2) explains how the sub-magnitudes are calculated. The first magnitude M1 represents the data split by only the first attribute, in our case medicine. It is calculated by taking the difference between the two results, medicine A and B and dividing it by the greater of these two. The division provides relative values in the range zero to one.
The second magnitude M2 represents the data split by the second partition attribute as well, in our case gender. It is calculated the same way as M1 with the difference that the sum of all partitions is taken and divided by their count m in the end.

Note that the denominator of any term can be zero. In order to avoid division by zero, if any of those terms is zero, the whole term is simply considered zero. Meaning if both Pr(G1) and Pr(G2) are zero the whole term $\frac{|\Pr(G1)-\Pr(G2)|}{\max(\Pr(G1),\Pr(G2))}$ is considered zero.

Applying this formula to the previous tables will yield the following results:

$M1 = \frac{|45.45\%-44.44\%|}{\max(45.45\%,44.44\%)} = \frac{1}{45} \approx 2.2\ \%$

	Cured	Not cured	Succes rate
Medicine A	5	6	45.45%
Medicine B	4	5	44.44%

$M2_1 = \left(\frac{|25\%-33\%|}{\max(25\%,33\%)}\right) = \frac{8}{33} \approx 24{,}2\ \%$

Gender: M	Cured	Not cured	Succes rate
Medicine A	1	3	25%
Medicine B	2	4	33%

$M2_2 = \left(\frac{|57\%-66\%|}{\max(57\%,66\%)}\right) = \frac{3}{22} \approx 13.63\ \%$

Gender: W	Cured	Not cured	Succes rate
Medicine A	4	3	57%
Medicine B	2	1	66%

$\Rightarrow M2 = \left(\frac{8}{33} + \frac{3}{22}\right) / 2 = \frac{25}{132} \approx 18.93\ \%$

Figure 21: Doctor patient example

This yields:

$M = \frac{1}{45} + \frac{25}{132} = \frac{419}{1980} \approx 21.16\ \%$

Interpreting the results the following observations can be made:
M1 is very low since medicine A surpasses medicine B only by very little. M2 on the other hand is larger since the difference between A and B is considerably greater. Therefore, the overall magnitude of 21.16 % can be seen as the result of a very narrow gap between A and B split by the first attribute thus making the occurrence not that significant and a bigger difference when split by the second attribute. The second time the difference is larger and makes the reversal therefore more important.

Examining the magnitude from another perspective shows the following: If a research concerning two different types of medicine shows, that A surpasses B by one percent when considering the aggregated data and B surpasses A by one percent when considering the disaggregated data the whole conclusion should be questioned. If the values are so close by each other the whole study might be set up in an insufficient way or the criteria in consideration are not fitting. Either way the occurrence of the paradox is not the crucial component because the whole study seems not very significant.

Using this additional value, all results found by the algorithm can be ranked and certain alerts can be set, so that reversals with a very low magnitude will not get as much attention as cases with a higher value.

5. Conclusion

Simpsonʾs paradox may still seem mysterious or unnatural whenever it is encountered. However, this is only a sensation that appears on first glance. The paradox itself always has a logical explanation as the examples in the beginning of this paper demonstrated.

Using the tools and techniques presented in this paper it is safe to say that Simpsonʾs paradox is resolved, as it can be detected and solved using the tools and techniques presented in this papier.

On one hand, it is possible to determine whether certain data holds a case of the paradox when grouped according to specific criteria. On the other hand, once found the paradox can be solved clearly determining which aggregation of data provides the correct answer.

However, finding this explanation and understanding how the different variables influence each other is the task of a sophisticated statistician. This means that although the paradox does not pose a threat to any survey it is important to thoroughly choose the necessary criteria and to understand the relationships in which the criteria in consideration are in.

6. References

- **Pearl Judea 2013.** Understanding Simpson's Paradox [Online] 2013. Technical Report R-414. University of California Los Angeles December 2013. http://ftp.cs.ucla.edu/pub/stat_ser/r414.pdf
- **Colonna Jean- François 2016.** [Online] 2013 – 2016. Franhttp://www.lactamme.polytechnique.fr/images/PARS.31.1_5.D/display.html
- **Schneiter Kady 2015.** [Online] Uta State University 2015 http://www.math.usu.edu/~schneit/CTIS/SP/
- **Martin Chris, Martin Alison 2015.** Simpson's Paradox: Why Smoking reduces the risk of cardiovascular disease [Online] 2015. http://www.crystallise.com/static/publications/PCV58%20Simpsons%20paradox%20smoking.pdf
- **Inglis-Arkell Esther 2013.** Simpson's Paradox "proves" smoking is good for you [Online] 2013. https://io9.gizmodo.com/simpsons-paradox-proves-smoking-is-good-for-you-1196099636
- **Liddell Mark 2016.** How statistics can be misleading [Online] 2016. TED-Ed https://ed.ted.com/lessons/how-statistics-can-be-misleading-mark-liddell
- **Gardner Martin 1976.** Mathematical Games [Online] n.D. Scientific American 03.1976. Page: 119 – 125 http://flowcytometry.sysbio.med.harvard.edu/files/flowcytometryhms/files/herzenbergfacshistory.pdf#129
- **Charig C. R., Webb D. R., Payne S.R., Wickham J. E. 1986.** Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy [Online] n.D. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1339981/
- **Simpson E. H. 1951.** The Interpretation of Interaction in Contingency Tables [Online] 2010. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 13, No. 2 http://www.epidemiology.ch/history/PDF%20bg/Simpson%20EH%201951%20the%20interpretation%20of%20interaction.pdf
- **Berman Steve, DalleMule Leandro, Greene Michael, Lucker John 2012.** Simpson's Paradox: a cautionary tale in advanced analytics [Online] 2012. Posted in: The Statistics Dictionary. https://www.statslife.org.uk/the-statistics-dictionary/2012-simpson-s-paradox-a-cautionary-tale-in-advanced-analytics
- **Freitas Alex A. 1998.** On objective measures of rule surprisingness [Online] 2006. Springer-Verlag. https://link.springer.com/chapter/10.1007/BFb0094799
- **Schmarzo Bill 2014.** Beware Simpson's Paradox [Online] 2014 https://infocus.emc.com/william_schmarzo/beware-simpsons-paradox/
- **Bickel P. J., Hammel E. A., O'Connell J.W. 1975.** Sex Bias in Graduate Admissions: Data from Berkeley [Online] 2008. Published by American

Association for the Advancement of Science http://www.unc.edu/~nielsen/soci708/cdocs/Berkeley_admissions_bias.pdf

- **Wikipedia.** Probabilistic Causation [Online] n.D. Last edited 30.12.2017. https://en.wikipedia.org/wiki/Probabilistic_causation
- **Pear Judea 2012.** The Do-Calculus Revisited [Online] 2012. Technical Report R-402. University of California Los Angeles August 2012. http://ftp.cs.ucla.edu/pub/stat_ser/r402.pdf
- **Wikipedia.** Simpson's paradox [Online] n.D. Last edited 27.12.2017. https://en.wikipedia.org/wiki/Simpson%27s_paradox
- **Yule, G. Udny 1903.** Notes on the Theory of Associations of Attributes in Statistics [Online] n.D. Published by Oxford University Press on behalf of Biometrika Trust. https://www.jstor.org/stable/2331677?seq=2#page_scan_tab_contents
- **SciShow 2017.** Statistical Paradoxes with MinutePhysics – SciShow Talk Show [Online] 2017. https://www.youtube.com/watch?v=FDsQB5Ug4SQ&feature=youtu.be
- **Yudkowsky Eliezer 2012.** Causal Diagrams and Causal Models [Online] 2012. http://lesswrong.com/lw/ev3/causal_diagrams_and_causal_models/
- **Aldrich John 1995.** Correlations Genuine and Spurious in Pearson and Yule [Online] n.D. Published in: Statistical Science 1995, Vol.10, No.4 364 – 376 http://www.economics.soton.ac.uk/staff/aldrich/spurious.PDF
- **Hitchcock Christopher 2010.** Probabilistic Causation [Online] 2010. Published in: Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/causation-probabilistic/
- **Edwards Anthony 2007.** A cautionary tale [Online] 2007. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2007.00223.x/full
- **Lehe Lewis, Powell Victor n.D.** Simpsonʾs Paradox [Online] n.D. http://vudlab.com/simpsons/
- **Porter Theodore M. 1998.** Karl Pearson [Online] 2017. https://www.britannica.com/biography/Karl-Pearson
- **Julious Steven A., Mullee Mark A. 1994.** Confounding and Simpsonʾs paradox [Online] n.D. http://www.bmj.com/content/309/6967/1480.full
- **Carlson Bruce W. 2016.** Simpsonʾs paradox [Online] 2016. https://www.britannica.com/topic/Simpsons-paradox
- **Freitas Alex A., Fabris Carem C. 2000.** Discovering Surprising Patterns by Detecting Occurences of Simpsonʾs Paradox [Online] n.D. Published by: Springer-Verlag London Limited 2000. https://link.springer.com/chapter/10.1007/978-1-4471-0745-3_10
- **Wright Bradley 2008.** Types of Causality [Online] 2008. http://www.everydaysociologyblog.com/2008/07/types-of-causal.html
- **Salkind Neil J. 2010.** Association Paradoxes [Online] n.D. Encyclopedia of research Design. Page: 1382 - 1383. https://books.google.de/books?id=HVmsxuaQl2oC&pg=PA1383&lpg=PA1383&

dq=pearson+paradox+skull&source=bl&ots=HQETHG_DnL&sig=CKzJnYlBAC9o7HQ0wmG9YpOIi1w&hl=en&sa=X&ved=0ahUKEwiBxoiX9NrYAhXR26QKHXrqA5AQ6AEIKTAA#v=onepage&q&f=false

- **Pearl, Judea 1995.** Causal diagrams for empirical research [Online] n.D. Technical report R-218-B. Published in: Biometrika December 1995. http://bayes.cs.ucla.edu/R218-B.pdf
- **Freitas Alex A., Fabris Carem C. 2006.** Discovering Surprising Instances of Simpson's Paradox in Hierarchical Multidimensional Data [Online] n.D. Published in: International Journal of Data Warehousing and Mining https://www.igi-global.com/article/international-journal-data-warehousing-mining/1762
- **Pearl Judea. 2009.** Causality, Models,Reasoning and Inference. Cambridge University Press; 2nd edition (September 14, 2009). New York
- **Delahaye Jean Paul. 2017**. Spektrum der Wissenschaft (02.2017). Page 70 -77. „Das Beunruhigende Paradox von Simpson. Spektrum der Wissenschaft Verlag, 2017. Heidelberg.

Note: All websites were last used by the author: 10.01.2018 between 09:00 and 18:00.

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, daß ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Ort	Datum	Unterschrift
Pinneberg	13.02.2018