Simpson’s paradox
Master Practical Course “Data Analysis with Python”
(WiSe 2016/17)

V. Stürzer, A. Rutabandama,
S. Grundner-Culemann, C. Kiesl, T. Bunk
(Gruppe TASC-Force)

Lehr- und Forschungseinheit für Programmier- und Modellierungssprachen
Institut für Informatik
Ludwig-Maximilians-Universität München

22. November 2016
Introduction

Veronika Stuerzer
Example

- There are only two schools in a certain district
- The graduation rate for girls in school A is higher than the graduation rate for boys in school A
- The graduation rate for girls in school B is higher than the graduation rate for boys in school B

Does it follow that the graduation rate for girls in this district is higher than the graduation rate for boys?
Does it follow that the graduation rate for girls in this district is higher than the graduation rate for boys?

a. Yes, girls have a higher graduation rate
b. No, boys have a higher graduation rate
c. No, boys and girls have an equal graduation rate
d. No, there is not enough information to answer this question
Answer

- The intuitive answer might be answer A, but depending on the ratio of boys and girls in each group, the result could be answer B or C as well ⇒ Answer D is correct.
- If boys actually had a higher graduation rate than girls, even though girls have a higher one in both schools, this effect is called the Simpson’s paradox.
- The paradox is caused when only the percentage, but not the ratio is given.
Simpson’s Paradox

- Simpson’s Paradox := a paradox in which a trend appears in different groups of data but disappears or reverses when these groups are combined
- often encountered in social-science and medical-science statistics
- described by British statistician Edward Hugh Simpson (*1922) in 1951
- similar effects have already been mentioned in 1899
Examples

Alexandre Rutabandama
<table>
<thead>
<tr>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ vote on the Civil Rights Act by chamber of congress and party</td>
</tr>
<tr>
<td>▪ A survey on women’s role in the 19th century vs 20th to date</td>
</tr>
</tbody>
</table>
Case description

- Almost 50 years ago in USA, Congress passed the landmark Civil Rights Act of 1964
- Senators were to vote for or against the act
- 152/248 democrats voted for the act making 61%
- 138/172 republicans voted for the act making 80%
- Conclusion: Republicans are more supportive of civil rights legislation
Simpson’s Paradox at work

Numbers don’t lie

- Voting patterns are broken down into Northern and Southern
- In the north 145/154 (94%) democrats voted for the act, against 138/162 (85%) republicans
- In the south 7/94 (7%) democrats voted for the act, against 0/10 (0%) republicans

**Conclusion:** Democrats are indeed more supportive of civil rights legislation
150 women were asked whether they were home mums. Answer: yes or no.

These women were asked at two different periods in time: 1750-1915 and 1915-2016.

Result/Conclusion: they were less home mums between 1750-1915 (49%) than 1915-2016 (55%).

<table>
<thead>
<tr>
<th></th>
<th>1750-1915</th>
<th>1915-2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home mums</td>
<td>74/150(49%)</td>
<td>82/150(55%)</td>
</tr>
</tbody>
</table>

Table 1: Percentage of women who were home mums between 1750-1915 and 1915-2016.
Simpson’s paradox at work

Once again numbers don’t lie!

- An additional question to find out whether they had a school education. Answer: yes or no
- They were more home mums with school education (34%) between 1750-1915 than (24%) between 1915-2016
- They were also more home mums with no school education between 1750-1915 (80%) than 1915-2016 (70%)
- This contradicts the conclusion in Table 1

<table>
<thead>
<tr>
<th></th>
<th>1750-1915</th>
<th>1915-2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>school-education</td>
<td>34/100 (34%)</td>
<td>12/50 (24%)</td>
</tr>
<tr>
<td>No-school-education</td>
<td>40/50 (80%)</td>
<td>70/100 (70%)</td>
</tr>
</tbody>
</table>

Table 2: Percentage of women who were home mums with or without a school education, between 1750-1915 and 1915-2016
The mathematical approach

Sophia Grundner-Culemann
Assume the following mathematical expressions:

- \( \frac{f_1}{F_1} > \frac{g_1}{G_1} \)
- \( \frac{f_2}{F_2} > \frac{g_2}{G_2} \)

Does it follow that \( \frac{f_1 + f_2}{F_1 + F_2} > \frac{g_1 + g_2}{G_1 + G_2} \)? \((*)\)

a) Yes, the first expression is always greater than the second.
b) No, the first expression is always smaller than the second.
c) No, the first and second expressions are always equal.
d) None of the above.
Answer

Prove: a, b and c are "sometimes" true, but not "always". (Thus, d holds.)

\[ a \ ">" : \frac{1}{1} > \frac{0}{1} \quad \implies \quad \frac{1+1}{1+1} = 1 > 0 = \frac{0+0}{1+1} \]

\[ b \ "<" : \frac{1}{50} > \frac{0}{1} \quad \implies \quad \frac{3}{53} < \frac{10}{31} \]

\[ c \ "=" : \frac{31}{96} > \frac{0}{1} \quad \implies \quad \frac{33}{99} = \frac{100}{300} \]
Why are we the fools of our intuition?

- \( \frac{f_1}{F_1} > \frac{g_1}{G_1} \)
- \( \frac{f_2}{F_2} > \frac{g_2}{G_2} \)

It does indeed follow that \( \frac{f_1}{F_1} + \frac{f_2}{F_2} > \frac{g_1}{G_1} + \frac{g_2}{G_2} \).
Why are we the fools of our intuition?

- \( \frac{f_1}{F_1} > \frac{g_1}{G_1} \)
- \( \frac{f_2}{F_2} > \frac{g_2}{G_2} \)

It does indeed follow that \( \frac{f_1}{F_1} + \frac{f_2}{F_2} > \frac{g_1}{G_1} + \frac{g_2}{G_2} \).

And when \( F_1 = F_2 \) and \( G_1 = G_2 \), our intuition is right:

\[
\frac{f_1 + f_2}{F_1 + F_1} = \frac{f_1 + f_2}{2F_1} = \frac{f_1}{2F_1} + \frac{f_2}{2F_1} = \frac{1}{2} \left( \frac{f_1}{F_1} + \frac{f_2}{F_2} \right) > \frac{1}{2} \left( \frac{g_1}{G_1} + \frac{g_2}{G_2} \right) = \ldots = \frac{g_1 + g_2}{G_1 + G_1}
\]
Why are we the fools of our intuition?

- \[ \frac{f_1}{F_1} > \frac{g_1}{G_1} \]
- \[ \frac{f_2}{F_2} > \frac{g_2}{G_2} \]

It does indeed follow that \[ \frac{f_1}{F_1} + \frac{f_2}{F_2} > \frac{g_1}{G_1} + \frac{g_2}{G_2} \).

And when \( F_1 = F_2 \) and \( G_1 = G_2 \), our intuition is right:

\[
\frac{f_1+f_2}{F_1+F_1} = \frac{f_1+f_2}{2*F_1} = \frac{f_1}{2*F_1} + \frac{f_2}{2*F_1} = \frac{1}{2} \left( \frac{f_1}{F_1} + \frac{f_2}{F_2} \right) > \frac{1}{2} \left( \frac{g_1}{G_1} + \frac{g_2}{G_2} \right) = \ldots = \frac{g_1+g_2}{G_1+G_1}
\]

But: in general \[ \frac{f_1}{F_1} + \frac{f_2}{F_2} = \frac{f_1*F_2+f_2*F_1}{F_1*F_2} \neq \frac{f_1+f_2}{F_1+F_2} \]

and therefore \[ \frac{f_1+f_2}{F_1+F_2} \neq \frac{g_1+g_2}{G_1+G_2} \]
Further mathematical considerations

Christoph Kiesl
The simpsons paradox can be interpreted in terms of vectors

\[
\frac{f_1}{F_1} \equiv (f_1, F_1) \in \mathbb{R}^2
\]

It follows that

\[
\frac{f_1 + f_2}{F_1 + F_2} \equiv (f_1 + f_2, F_1 + F_2) = (f_1, F_1) + (f_2, F_2)
\]

The value of the fraction is equal to the steepness of the vector.

A case of Simpson’s paradox occurs if \( A_1, A_2 \in \mathbb{R}^2 \) are steeper than \( B_1, B_2 \in \mathbb{R}^2 \) respectively, but \( B_1 + B_2 \) is steeper than \( A_1 + A_2 \).
Vector Interpretation - Example
Stochastical analysis I

- How often does Simpson’s paradox occur?
- This depends on the number of categories
- Experiments show that the probability is smaller than 2% for two categories and rapidly decreases.
Stochastical analysis II
How to avoid the Simpson’s Paradox

Thomas Bunk
How to avoid the Simpson’s Paradox

There is no single mathematical property that all instances of SP have in common

- Thus, there exists no universal rule to prevent cases of SP
How to prevent the occurrence of the SP in our previous examples? (1/2)

In the subsets of the aggregated data the ratio has to be respected:

<table>
<thead>
<tr>
<th></th>
<th>1750-1915</th>
<th>1915-2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>School-education</td>
<td>34/100 (34%)</td>
<td>12/50 (24%)</td>
</tr>
<tr>
<td>No school-education</td>
<td>40/50 (80%)</td>
<td>70/100 (70%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1750-1915</th>
<th>1915-2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housewives</td>
<td>74/150 (49%)</td>
<td>82/150 (55%)</td>
</tr>
</tbody>
</table>
How to prevent the occurrence of the SP in our previous examples? (1/2)

In the subsets of the aggregated data the ratio has to be respected:

<table>
<thead>
<tr>
<th></th>
<th>1750-1915</th>
<th>1915-2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>School-education (old)</td>
<td>34/100 (34%)</td>
<td>12/50 (24%)</td>
</tr>
<tr>
<td>School-education (new)</td>
<td>17/50 (34%)</td>
<td>12/50 (24%)</td>
</tr>
<tr>
<td>No school-education (old)</td>
<td>40/50 (80%)</td>
<td>70/100 (70%)</td>
</tr>
<tr>
<td>No school-education (new)</td>
<td>40/50 (80%)</td>
<td>35/50 (70%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1750-1915</th>
<th>1915-2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housewives (old)</td>
<td>74/150 (49%)</td>
<td>82/150 (55%)</td>
</tr>
<tr>
<td>Housewives (new)</td>
<td>57/100 (57%)</td>
<td>47/100 (47%)</td>
</tr>
</tbody>
</table>
How to prevent the occurrence of the SP in our previous examples? (2/2)

- If the same ratio of the subsets is taken into account, the paradoxum never appears in the first place.
- Both calculations (i.e. the subsets and the aggregation) are statistically accurate.
- In our shown examples the results were computed correctly, though the wrong conclusions have been drawn.
The context is important!

Any statistical relationship between two variables can be reversed by adding an additional factor.

- Without any further context, no answer can be given about what factors should be included in the data analysis.
How can the previous examples be interpreted? (1/2)

Let’s have a look at the party voting example:

<table>
<thead>
<tr>
<th></th>
<th>Democrat</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern</td>
<td>94% (145/154)</td>
<td>85% (138/162)</td>
</tr>
<tr>
<td>Southern</td>
<td>7% (7/94)</td>
<td>0% (0/10)</td>
</tr>
<tr>
<td>Total</td>
<td>61% (152/248)</td>
<td>80% (138/172)</td>
</tr>
</tbody>
</table>

- How can the democrats be more supportive in each region (north and south) compared to the republicans, yet in aggregation they are less supportive?
  - Verify that there is a logical reasoning behind the emerged differences.
How can the previous examples be interpreted? (2/2)

Let’s have a look at the party voting example:

<table>
<thead>
<tr>
<th></th>
<th>Democrat</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern</td>
<td>94% (145/154)</td>
<td>85% (138/162)</td>
</tr>
<tr>
<td>Southern</td>
<td>7% (7/94)</td>
<td>0% (0/10)</td>
</tr>
<tr>
<td>Total</td>
<td>61% (152/248)</td>
<td>80% (138/172)</td>
</tr>
</tbody>
</table>

- Hypothetical question: How can the study be interpreted if the regions (north, south) are replaced by cat- and dog-owners for instance?
Conclusion

- Don’t blindly trust available statistics
- Lots of half-truth statements can be created this way, only to be abused for justifying one’s actions
References

Statistics That Lead Us Astray

The logic of Simpson’s paradox

Vector Interpretation
Simpson’s paradox in psychological science

Von Zahlen geblendet

Simpson-paradox-statistik