# An Exploration in the Intervals of Simpson's Paradox

Presented by: Raymond Gontkovsky and Julian Sobieski
Faculty Advisor: Dr. Darci Kracht

## Abstract

There exists specific scenarios in which two individual entities are evaluated and compared in two categories and one entity maintains a higher average in both given categories, while the other entity maintains a higher average overall; this occurrence is known as Simpson's Paradox. The discrepancy between the intuitive understanding of averaging averages and the correct method in adding averages leads to the paradoxical nature of Simpson's Paradox. With our project, we aspire to identify which conditions must be present for Simpson's Paradox to occur. First, we will explain an applied example of Simpson's Paradox. Second, we will define a model for Simpson's Paradox. From there, we will present our research in classifying interval restrictions which allow Simpson's Paradox to occur or prevent it from occurring entirely. Finally, we will present our research in further classifying Simpson's Paradox as a study of relationships and ratios.

## Introduction and Real Example

Simpson's Paradox occurs in various statistical settings, including but not limited to sports. In basketball, we can compare two players on the basis of their two point, three point, and overall field goal percentages. If one player maintains a higher percentage in both two point and three point averages while the other player maintains a higher percentage in the average of combined field goals, Simpson's paradox is yielded. In the case of Trevor Huffman and Bryan Bedford, Huffman maintained a higher percentage in the individual categories, while Bedford maintained a higher percentage in combined field goals.

Kent State Men's Basketball: 2000-2001 Conference Games Only (18 games)[1]

| | Trevor Huffman | | | Bryan Bedford | | |
|---|---|---|---|---|---|---|
| | Made | Attempts | Average | Made | Attempts | Average |
| Two-pointers | 57 | 127 | **0.449** | 13 | 30 | 0.433 |
| Three-pointers | 35 | 100 | **0.350** | 0 | 1 | 0.000 |
| All field goals | 92 | 227 | 0.405 | 13 | 31 | **0.419** |

To depict Simpson's Paradox as a model, the data above can be represented by the notation of Table 1 found in Elementary Analysis.

## Elementary Analysis

Table 1:

| | Player $X_{(x,t)}$ | Player $Y_{(y,s)}$ |
|---|---|---|
| **2 Points** | $X_2 = \frac{x_2}{t_2}$ | $Y_2 = \frac{y_2}{s_2}$ |
| **3 Points** | $X_3 = \frac{x_3}{t_3}$ | $Y_3 = \frac{y_3}{s_3}$ |
| **Totals** | $R_X = \frac{x_2 + x_3}{t_2 + t_3}$ | $R_Y = \frac{y_2 + y_3}{s_2 + s_3}$ |

Where $X_2, X_3, Y_2, Y_3, R_X, R_Y$ are percentages. Simpson's Paradox states that while $X_2 > Y_2$ and $X_3 > Y_3$, $R_Y$ is greater than $R_X$.

Prove: $X_3 < R_X < X_2$

*Assume* $X_3 < X_2$:

$$\frac{x_3}{t_3} < \frac{x_2}{t_2}$$
$$x_3 t_2 < x_2 t_3$$

Part 1:

$$x_3 t_2 + x_3 t_3 < x_2 t_3 + x_3 t_3$$
$$x_3(t_2 + t_3) < t_3(x_2 + x_3)$$

$X_3 = \frac{x_3}{t_3} < \frac{x_2 + x_3}{t_2 + t_3} = R_X \rightarrow R_X > X_3$

Part 2:

$$x_3 t_2 + x_2 t_2 < x_2 t_3 + x_2 t_2$$
$$t_2(x_3 + x_2) < x_2(t_2 + t_3)$$

$R_X = \frac{(x_3 + x_2)}{(t_2 + t_3)} < \frac{x_2}{t_2} = X_2 \rightarrow R_X < X_2$

For $X_3 < X_2$, as $R_X > X_3$ and $R_X < X_2$, $\boldsymbol{X_3 < R_X < X_2}$ and player X's overall must be between anywhere between the averages of the two categories. Likewise, the same holds true for player Y on order of the same operation.

## Case II

Given: $X_2 > X_3$, $Y_2 > Y_3$, $X_2 > Y_2$, and $X_3 > Y_3$

Prove: $Y_2 > X_3$ can yield Simpson's Paradox as $R_Y$ can be > $R_X$

Assume: $s_3 = 1$, $y_3 = 0$

For an $s_2 >> 1 = s_3$, $\boldsymbol{R_Y} = \frac{y_2 + y_3}{s_2 + s_3} = \frac{y_2}{s_{2+1}} \approx \frac{y_2}{s_2} = \boldsymbol{Y_2}$

Assume: $t_2 = 1$, $x_2 = 1$

For a $t_3 \gg 1 = t_2$, $\boldsymbol{R_X} = \frac{x_2 + x_3}{t_2 + t_3} = \frac{x_3 + 1}{t_3 + 1} \approx \frac{x_3 + 1}{t_3}$

$\rightarrow X_3 + \frac{1}{t_3}$; $\frac{1}{t_3}$ is negligible as $t_3 \gg 1$, so $X_3 + \frac{1}{t_3} \approx \boldsymbol{X_3}$

The approximation is consistent considering $t_2, t_3, s_2, s_3,$ are all non-zero and $R_i \neq i_{2,3}$, where $i \in \{X, Y\}$

As $\boldsymbol{R_Y} \approx \boldsymbol{Y_2} > \boldsymbol{X_3} \approx \boldsymbol{R_X}$, Simpson's Paradox is yielded.

## Case I

Given Simpson's Paradox necessitates $X_2 > Y_2$ and $X_3 > Y_3$, and the situation assumes $X_2 > X_3$ and $Y_2 > Y_3$, prove X and Y's intervals do not overlap and Simpson's Paradox cannot occur for $X_3 > Y_2$.

*Proof:* Each subject's overall average ranges between two possible averages, as is evident in Elementary Analysis proof.

When calculating $R_X$ and $R_Y$ which are the average scores for X and Y we use the formulas:

$R_X = \frac{x_2 + x_3}{t_2 + t_3}$ and $R_Y = \frac{y_2 + y_3}{s_2 + s_3}$, where t and s denote the number of total attempts for their respective categories and x and y denote the number of success in each respective category.

The minimum for $R_X \approx X_3$ and the maximum value for $R_Y \approx Y_2$ as $X_3 < R_X < X_2$ and $Y_3 < R_Y < Y_2$. While $R_X > X_3$, $R_y < Y_2$, and $X_3 > Y_2$ there cannot be overlap between the intervals of possible averages for $R_X$ & $R_Y$ as $R_X > X_3 > Y_2 > R_Y$. No matter how weighted the categories are, $\boldsymbol{R_X > R_Y}$.

## Research

| $\Theta_X = \frac{X_2}{X_3}$ | $\Theta_Y = \frac{Y_2}{Y_3}$ | $\Omega_X = \frac{t_2}{t_3}$ | $\Omega_Y = \frac{s_2}{s_3}$ |
|---|---|---|---|
| $\Omega_R = \frac{\Omega_X}{\Omega_Y}$ | Equation 1: $R_X = X_3 \frac{\Theta_X \Omega_X + 1}{\Omega_X + 1}$ | Equation 2: $R_Y = Y_2 \frac{1 + \frac{1}{\Theta_Y \Omega_Y}}{1 + \frac{1}{\Omega_Y}}$ | |

## Conclusion

Using these equations, we can evaluate the overall averages of either player as an expression of proportions. Further research of these relationships allows evaluation of Simpson's Paradox as a continuous condition rather than a discreet condition. One possible analysis involves the evaluation of equations 1 and 2 as a limit of any of the newly defined variables. Analysis of Simpson's Paradox as a continuous condition is the next step in applying Simpson's Paradox to new models.

Prove: $R_X > R_Y$ for $\Omega_R = 1$

$$X_2 \frac{1 + \frac{1}{\Theta_X \Omega_X}}{1 + \frac{1}{\Omega_X}} > Y_2 \frac{1 + \frac{1}{\Theta_Y \Omega_Y}}{1 + \frac{1}{\Omega_Y}}$$

Since $\Omega_X = \Omega_Y$, $\frac{X_2}{Y_2} \frac{1 + \frac{1}{\Theta_X \Omega_X}}{1 + \frac{1}{\Omega_X}} > \frac{1 + \frac{1}{\Theta_Y \Omega_X}}{1 + \frac{1}{\Omega_X}}$

$$\rightarrow \boldsymbol{\frac{X_2}{Y_2} \frac{1 + \frac{1}{\Theta_X \Omega_X}}{1} > \frac{1 + \frac{1}{\Theta_Y \Omega_X}}{1}}$$

As $\Theta_Y > \Theta_X$ and $X_2 > Y_2$, the inequality stands true and Simpson's Paradox cannot be yielded for $\Omega_R = 1$.

## Contact

< Raymond Gontkovsky and Julian Sobieski>
<Kent State University; Choose Ohio First>
Email: rgontkov@kent.edu, jsobiesk@kent.edu
Phone: 330-235-3882, 330-861-2059

## References

1. "Advances in Recreational Mathematics," by Darci L. Kracht: Department of Mathematical Sciences, Kent State University. 2003.