STA 6384, Report 2.15

Carson Slater *Baylor University*

Problem: Work Agresti's problem 2.21, p. 66.

Smith and Jones are baseball players. Smith has a higher batting average than Jones in each of K years. Is is possible that for the combined data from the K years, Jones has the higher batting average? Explain, creating some data with $K = 2$ to illustrate.

Yes, it's entirely possible. This situation is a classic example of **Simpson's Paradox**, where a trend that appears in different groups of data disappears or reverses when these groups are combined.

We can demonstrate this with the real-world batting statistics of David Justice and Derek Jeter from the 1995 and 1996 seasons. For this example, **Smith = David Justice** and **Jones = Derek Jeter**.

Data by Individual Year

First, let's look at contingency tables for each year, showing hits versus outs (No Hit).

1995 Season

In 1995, Justice had the higher batting average.

Player	Hit	No Hit	Total At-Bats	Batting Average
David Justice	104	307	411	**0.253**
Derek Jeter	12	36	48	0.250

1996 Season

In 1996, Justice again had the higher batting average.

Player	Hit	No Hit	Total At-Bats	Batting Average
David Justice	45	95	140	**0.321**
Derek Jeter	183	399	582	0.314

So, in each individual year ($K = 2$), Smith (Justice) had a higher batting average than Jones (Jeter).

Combined Data

Now, we create a single contingency table by combining the data from both years.

Player	**Hit**	**No Hit**	**Total At-Bats**	**Batting Average**
David Justice	$104 + 45 = 149$	$307 + 95 = 402$	551	0.270
Derek Jeter	$12 + 183 = 195$	$36 + 399 = 435$	630	**0.310**

When the data is combined, the trend reverses. Derek Jeter (Jones) has a significantly higher overall batting average (0.310) than David Justice (Smith) (0.270).