22.8: Beware of Simpson's Paradox

The contingency tables presented above represent summaries of large numbers of observations, but summaries can sometimes be misleading. Let's take an example from baseball. The table below shows the batting data (hits/at bats and batting average) for Derek Jeter and David Justice over the years 1995-1997:

Player	1995		1996		1997		Combined	
Derek Jeter	12/48	.250	183/582	.314	190/654	.291	385/1284	**.300**
David Justice	104/411	**.253**	45/140	**.321**	163/495	**.329**	312/1046	.298

If you look closely, you will see that something odd is going on: In each individual year Justice had a higher batting average than Jeter, but when we combine the data across all three years, Jeter's average is actually higher than Justice's! This is an example of a phenomenon known as *Simpson's paradox*, in which a pattern that is present in a combined dataset may not be present in any of the subsets of the data. This occurs when there is another variable that may be changing across the different subsets – in this case, the number of at-bats varies across years, with Justice batting many more times in 1995 (when batting averages were low). We refer to this as a *lurking variable*, and it's always important to be attentive to such variables whenever one examines categorical data.