Simpson’s Paradox

- In 1951 E H Simpson published a seminal result in statistics which every Data Miner needs to be aware of (although lots aren’t!)
- His result is called a paradox because of the situation it leaves us in
- It arises from an easily understandable property of simple fractions

An Example of Simpson’s Paradox

- Simpson’s original scenario featured a baby mucking up a deck of cards but the phenomenon had been reported in a more serious form in 1934 relating to a 1910 study on tuberculosis in the USA
- The death rate for African Americans was shown to be statistically *lower* in Richmond than in New York
- The death rate for Caucasians was also statistically *lower* in Richmond than in New York
- What would you conclude about the combined death rate in Richmond compared to New York?

Example of Simpson’s Paradox II

- You’ve probably guessed what the statistics said ...
- The death rate for the total combined population of African Americans and Caucasians was *higher* in Richmond than in New York
- What’s going on?

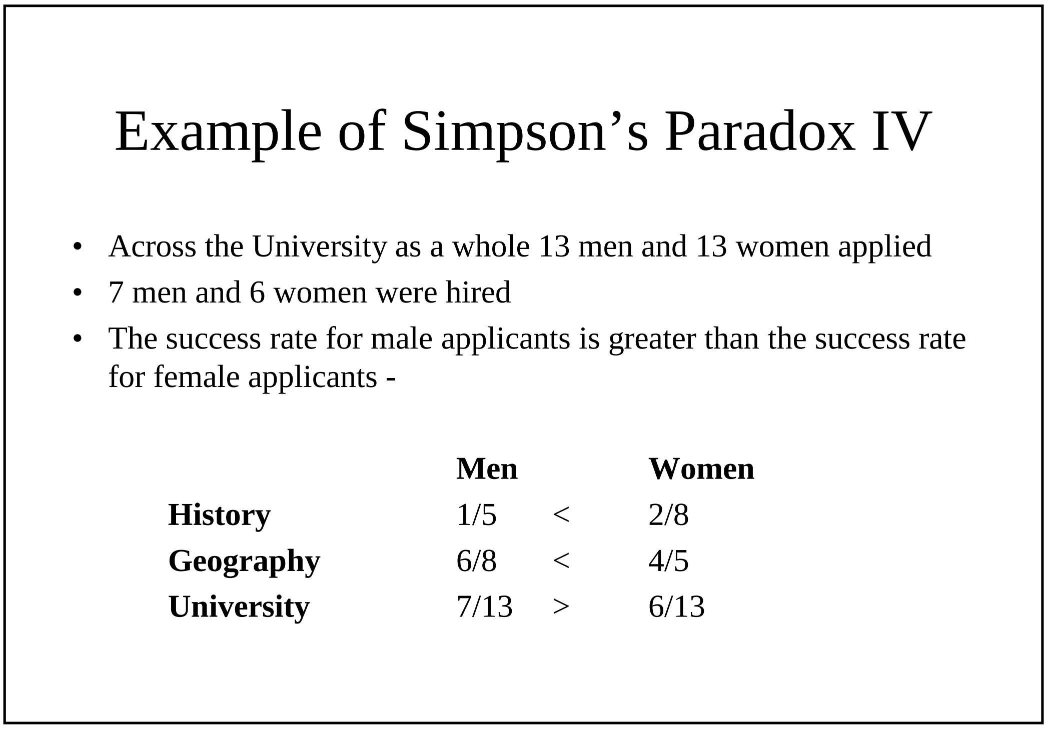
$$a/b < A/B$$

$$c/d < C/D$$

$$(a + c)/(b + d) > (A + C)/(B + D)$$

Example of Simpson’s Paradox III

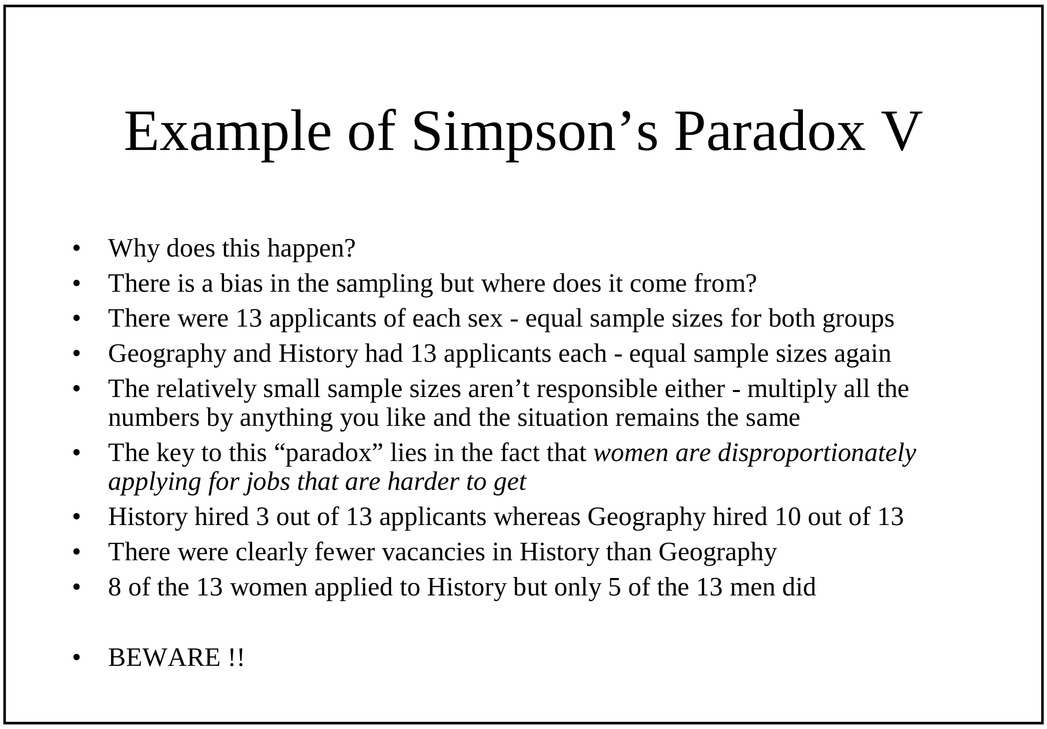
- Here’s a more contrived example which makes it easier for us to see what’s happening
- A university has vacancies in the departments of History and Geography and wishes to discriminate in favour of women
- In the History department
  - 5 men apply and 1 is hired
  - 8 women apply and 2 are hired
  - The success rate for men is 20% and for women it is 25%
  - The History department has favoured women over men
- In the Geography department
  - 8 men apply and 6 are hired,
  - 5 women apply and 4 are hired
  - The success rate for men is 75% and for women it is 80%
  - The Geography department has favoured women over men



Example of Simpson’s Paradox IV

- Across the University as a whole 13 men and 13 women applied
- 7 men and 6 women were hired
- The success rate for male applicants is greater than the success rate for female applicants -

	Men		**Women**
History	1/5	<	2/8
Geography	6/8	<	4/5
University	7/13	>	6/13



Example of Simpson’s Paradox V

- Why does this happen?
- There is a bias in the sampling but where does it come from?
- There were 13 applicants of each sex - equal sample sizes for both groups
- Geography and History had 13 applicants each - equal sample sizes again
- The relatively small sample sizes aren’t responsible either - multiply all the numbers by anything you like and the situation remains the same
- The key to this “paradox” lies in the fact that *women are disproportionately applying for jobs that are harder to get*
- History hired 3 out of 13 applicants whereas Geography hired 10 out of 13
- There were clearly fewer vacancies in History than Geography
- 8 of the 13 women applied to History but only 5 of the 13 men did

- BEWARE !!