On Aggregation

Simpson's Paradox

- When we aggregate data, we can have somewhat confusing results.

Consider two locations:

- Location 1 - Applied Accepted Rate

Male	100	80	80%
Female	16	14	87.5%

- Location 2 - Applied Accepted Rate

Male	36	18	50%
Female	120	76	63.3%

- Note that in each location, females are accepted at a higher rate than males (87.5% to 80% and 63.3% to 50%)

- But, if we combine the two locations, we find, somehow, the opposite:

Combined	Applied	Accepted	Rate
Male	136	98	72.1%
Female	136	90	66.2%

- So, is there discrimination against females occurring here? If so, how could we address it? Clearly we cannot "tell each location" to fix the problem . . . They each seem to be doing OK . . .

Another view

- Consider a new drug being tested among two groups of people. The combined results of the tests:

Overall	Number	Improved	Rate
Treated	136	98	72.1%
Placebo	136	90	66.2%

Now, separate by age

- Under 55 - Number Improved Rate

Treated	100	80	80%
Placebo	16	14	87.5%

- 55+ - Number Improved Rate

Treated	36	18	50%
Placebo	120	76	63.3%

- So, if you are just a person, taking the drug helps (over placebo) by 72.1% to 66.2%
- But, if you are under 55, the drug hurts (compared with placebo) by 80% to 87.5%
- And, if you are 55 or older, the drug hurts (compared with placebo) by 50% to 63.3% . . .

What should you do?

- (Apparently, forget your age, and just be a person :-)