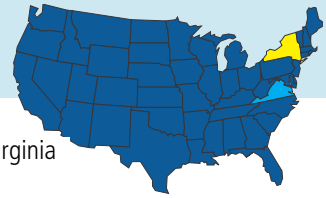
SIMPSON'S PARADOX



States of New York and Virginia

par·a·dox [par-uh-doks] –noun
1. a statement or proposition that seems self-contradictory or absurd but in reality expresses a possible truth.
2. a self-contradictory and false proposition.
3. any person, thing, or situation exhibiting an apparently contradictory nature.
4. an opinion or statement contrary to commonly accepted opinion.

(http://dictionary.reference.com/browse/paradox, April 14 2010)

In statistics about deaths caused by tuberculosis (TB) in New York and Richmond (Virginia) in 1910 the figures are very different for black and white people.
(Cohen and Nagel: "An Introduction to Logic and Scientific Method", 1934)

number of	population		deaths caused by TB	
	New York	Richmond	New York	Richmond
white people	4675174	80895	8365	131
black people	91709	46733	513	155
total	4766883	127628	8878	286

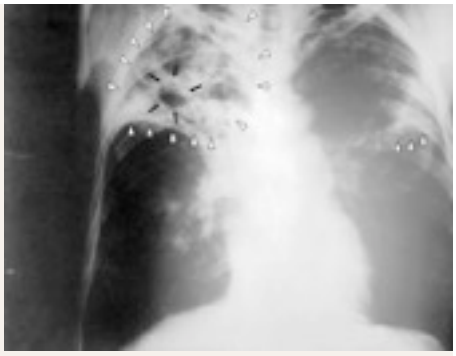
To calculate the relative frequencies the number of deaths is divided by the population number, e.g.

the relative frequency of deaths of white people in New York $= \frac{8{,}365}{4{,}675{,}174} \approx 0.0018$

relative frequency of deaths caused by TB		
	New York	Richmond
white people	0.0018	0.0016
black people	0.0056	0.0033
total	0.0019	0.0022

These relative frequencies now show a strange pattern:
If you want to live long and you are **white**, you should go to **Richmond**.
If you want to live long and you are **black**, you should go to **Richmond** as well.
But if you want to live long and you **do not care about the colour of your skin**, you should go to **New York**!

Tuberculosis or **TB** (short for tubercles bacillus) is a common and often deadly infectious disease caused by various strains of bacteria. Tuberculosis usually attacks the lungs but can also affect other parts of the body. It is spread mostly through the air, when people who have the disease cough, sneeze, or spit (but can also be passed on by milk of infected cows for example).
Most infections in humans result in asymptomatic, latent infection, and about one in ten latent infections eventually progresses to active disease, which, if left untreated, kills more than 50 % of its victims.



Chest X-ray of a patient with far-advanced tuberculosis

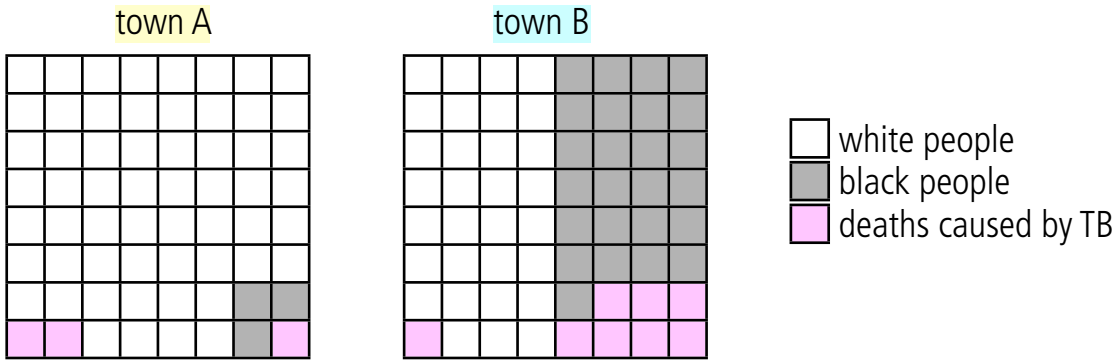
In the 17th and 18th century TB was one of the most feared epidemics. Until World War I it was death cause number one in Europe.
The classic symptoms are a chronic cough with blood-tinged sputum, fever and weight loss. (...) Treatment is difficult and requires long courses of multiple antibiotics. (...) Antibiotic resistance is a growing problem (...).
A third of the world's population are thought to be infected with M. tuberculosis, and new infections occur at a rate of about one per second. (...)
(http://en.wikipedia.org/wiki/Tuberculosis, April 14 2010)

For black and white people alone the chance to die of TB are higher in New York but for the total of the inhabitants the chance in Richmond is higher! This phenomenon is called **"Simpson's Paradox"** How can such a thing happen?

The statisticians Edward H. Simpson described this phenomenon in a technical paper in 1951, but others like Cohen and Nagel had mentioned this effect earlier.

It is easier to grasp the idea with smaller figures:

number of	population		deaths caused by TB	
	town A	town B	town A	town B
white people	60	32	2	1
black people	4	32	1	7
total	64	64	3	8



Again with the relative frequencies the same strange pattern appears:

relative frequency of deaths caused by TB		
	town A	town B
white people	0.0333	0.0313
black people	0.25	0.2188
total	0.0467	0.125

For a white person it is healthier to live in town B (with respect to TB).
For a black person it is healthier in town B as well.
But for a person who does not care about the colour of the skin it is healthier to live in town A!

The reason for this strange result can not be the size of the towns, as they are equally big. So it must be the **different composition of the population**: in town A there are only 4 black people, in town B there are 32! The minority makes the difference.

In this example the black people form a small minority in town A.
1 death out of 4 black people causes a high rate within the minority but has little effect on the total rate.
In town B black and white people balance each other.
7 deaths out of 32 black people cause a lower rate within the black population than in town A but have a big influence on the rate of the total population.

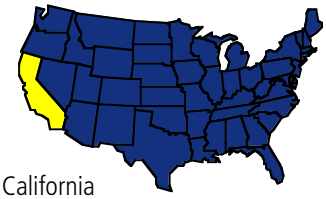
The death rate of the white people is nearly the same in both towns: 2 out of 60 and 1 out of 32. The rate of town A is higher but both figures have little influence on the total rate.

Simpson's paradox can appear when the groups compared are differently composed!

1

BERKELEY SEX BIAS CASE

David Freedman, Robert Pisani and Roger Purves. "Statistics" (3rd edition). W.W. Norton, 1998, p. 19.



State of California

One of the best known real life examples of Simpson's paradox occurred when the University of Berkeley was sued for bias against women who had applied for admission to graduate schools there. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.

	applicants	**admitted**
men	2590	46 %
women	1835	30 %

However when examining the individual departments, it was found that no department was significantly biased against women. In fact, most departments had a "small but statistically significant bias in favour of women".

department	**men**		**women**	
	applicants	**admitted**	**applicants**	**admitted**
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Show that Simpson's paradox occurs and explain why.

2

BATTING AVERAGES

Ken Ross. "A Mathematician at the Ballpark: Odds and Probabilities for Baseball Fans" Pi Press, 2004

A common example of Simpson's Paradox involves the batting averages of players in professional baseball. It is possible for one player to have a higher batting average (e.g. 12/48 = 0.25 runs per bat) than another player during a given year, and to do so again during the next year, but he has a lower batting average when the two years are combined. This phenomenon can occur when there are large differences in the number of at-bats between the years. (The same situation applies to calculating batting averages for the first half of the baseball season, and during the second half, and then combining all of the data for the season's batting average.)

A real-life example is provided by Ken Ross and involves the batting average of two baseball players, Derek Jeter and David Justice, during the baseball years 1995 and 1996:

	1995	**1996**	**combined**
Derek Jeter	12/48	183/582	195/630
David Justice	104/411	45/140	149/551

Show that Simpson's paradox occurs and explain