1 How to Lie with Statistics

"There are three kinds of lies: lies, damned lies, and statistics", as Mark Twain is reputed to have said. In this lecture we will see why statistics are so easy to misuse, and some of the ways we must be careful while evaluating statistical claims.

First a very simple but important point about statistics. Statistics cannot be used to establish causation, they can only show correlations. As an example, the governor of a certain state is concerned about the test scores of high school students in the state. One of his aides brings up an interesting statistic: there is a very strong link between student test scores and the taxes paid by the parents of the student. The parents of high scoring students pay more taxes. The aide's suggestion for increasing student test scores is unusual; sharply increase tax rates. Surely student test scores will follow! The fallacy, of course, is that even though there is a correlation between test scores and parents taxes, there is likely no causal connection. A better explanation is that there is some hidden variable that explains the correlation. In this case the obvious choice is the income of the parents. This determines the taxes paid. And since the quantity of high school that a student attends is to a large extent determined by the parent's income, we see a caual link from parent's income to both taxes and test scores.

Let us now turn to a very important paradox in probability called Simpson's paradox, described by Simpson in his 1951 paper. Let us start with an example, which studies the 20 year survival rate of smokers. A paper by Appleton, French, and Vanderpump (1996, *American Statistician*) surveyed 1314 English woman in 1972-74 and after 20 years. The first table they got is

Smoker	Dead	Alive	Total	% Dead
Yes	139	443	582	24
No	230	502	732	31
Total	369	945	1314	28

From this data one might conclude that non-smoking kills. How does one explain this unusual data. The answer lies in the composition of the two groups. Smoking was unpopular in the middle of 20th century among women in England, and it increased only in the 70's but it was mostly young woman who started to smoke. Therefore the sample of smokers in the study was heavily biased towards young women, whose expected lifespan was much larger than 20 years. This becomes more clear when we look more closely at the results of the study broken down by age group:

Age group	18-24		25-34		35-44		45-54		55-54	
Smoker	Y	N	Y	N	Y	N	Y	N	Y	N
Dead	2	1	3	5	11	7	27	12	51	40
Alive	53	61	121	152	95	114	103	66	64	81
Ratio	2.3		0.75		2.4		1.44		1.61	

The interesting thing to notice is that the fatality rates are significantly higher for smokers in almost every age group. The data could be made even more dramatic by increasing the smoking fatalities in the couple of exceptional groups by one or two, thereby achieving the following strange result: in each separate category, the percentage of fatalities among smokers is higher, and yet the overall percentage of fatalities among smokers is lower. This is an example of so called **Simpson's paradox**.

Wikipedia gives a good explanation of Simpson's paradox by another example.

A real word example is the Berkeley sex bias case. In 1973 U.C. Berkeley was sued for bias against woman applying to grad school. Data showed that 44% of men were admitted and only 30% of women. Since admittance is decided by departments, they started to investigate which department is "discriminating" women. It turned out that none of them did! Here is an admittance data for largest departments:

Department	#male applicants	#female applicants	%male admit	%female admit
A	825	108	62	82
B	560	25	63	68
C	325	593	37	34
D	417	375	33	54

The explanation is that women applied in larger numbers to departments that had lower admittance rates.

2 Zipf's Law

Zipf's law was published by Harvard linguist George Kingsley Zipf in 1949. He observed that the most frequently-occurring word was "the", that accounts 7% of all words occupancies. The second most common was "of", that accounts 3.5%. The third most common word was "and" that accounts 2.2%. The i-th most common word accounted about $\frac{1}{i} \cdot 7\%$.

Similar "law", where the frequency of an item is inversely proportional to its rank in the frequency table, is commonly observed in many kind of phenomena, including

- Frequency of accesses to web pages
- Frequency of keyword usage in a search engine
- Population of cities
- Income distribution among 3% of richest individuals
- Frequency of web page pointers
- The degree of nodes of Internet graph

This gives so called Zipf's distribution, where the probability of k-th largest item is

$$\Pr[f(k)] \approx \frac{1}{k^s},$$

for some s, typically $s = 1$ or a little bit larger.