There are three kinds of lies: lies, damned lies, and statistics.

In daily life, we often analyze a set of data as a whole, but that’s may lead to a problem.

In a university, some female students argue that they suffered gender bias, because women applying are significantly less likely than men to be admitted. However, when examining the individual departments, it appears that women applying are more likely to be admitted than men.

This is called Simpson’s paradox, which means for a set of data, the trend of each part is different from the overall trend.

However, how can that be?

Basically, the reason is lurking variable, which is a variable that is not included as an explanatory or response variable in the analysis, but can affect the interpretation of relationships between variables.

For example, in this case, the lurking variable is the admission difficulties of different departments. And women tended to apply to more competitive Business School, while men tended to apply to less competitive Law School.

We can represent each acceptance rate as a vector. The horizontal axis represents the total number of applications, and the vertical axis represents the number of accepted students. The slope of each vector is the corresponding acceptance rate. The graph shows that the overall acceptance rate depends on the sum of individual vectors.

What will happen if we change the number of applicants while maintaining the same acceptance rate?

If the number of women applicants of Law School increased to three times, the overall acceptance rate of female students will be higher.

Let’s see another example.

In this example, the relationship between an applicant’s GPA and the rejected rate by a university.

Each dot represents an applicant, and the horizontal and vertical axis show the corresponding GPA and the rate to be rejected.

If we only analyze the overall trend, we may conclude that a higher GPA leads to a higher rejected rate.

However, these data come from four different departments, and the admission difficulties of each of them are different.

Let’s color the dots and analyze the data by department.

Now, in each department, a higher GPA leads to a lower rejected rate.

In this case, the competitiveness of each department is a lurking variable. A more competitive department tend to have applicants with high GPA, and higher rejected rate at the same time.

If we ignore the lurking variable and analyze the data as a whole, we will get the wrong result.

Similar situations can happen in other cases, for example, if the horizontal and vertical axis represent the dose of a drug and the response. The lurking variable can be the differences in dose between different groups of people.

Now, you know how easy people can be tricked by Statistics. Therefore, to avoid statistical paradoxes, like Simpson’s paradox, it is very important to be aware that they are possible, and we need more detailed information to understand the actual meaning behind the data.