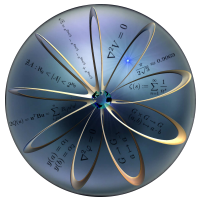

planetmath.org

Math for the people, by the people.

Simpson's paradox

Canonical name	SimpsonsParadox
Date of creation	2013-03-22 14:42:05
Last modified on	2013-03-22 14:42:05
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	8
Author	CWoo (3771)
Entry type	Definition
Classification	msc 62H17

Before describing what a *Simpson's paradox* is, let's start with a hypothetical example. During a particular summer, an experiment was conducted to find out the preference between two types of beverages: soda and lemonade. The data was drawn from two locations: city and rural. In each location, the gender and the choice of drinks were collected. The results are summarized as follows:

location	gender	lemonade	soda	total	% preferring lemonade	http://planetmath.org/C
city	female	150	300	450	24.9%	1.10
	male	300	660	960	23.1%	
rural	female	285	860	1145	33.3%	1.10
	male	30	100	130	31.3%	

The odds ratio given that location = city is about 1.1, showing that females are about 10% more likely to drink lemonade than males. Because the conditional odds ratio given that location = rural is also 1.1, the same conclusion can be drawn.

Next, combine the results from both locations and form the following 2 by 2 contingency table:

gender	lemonade	soda	total	% preferring lemonade	odds ratio
female	435	1160	1595	27.3%	0.86
male	330	760	1090	30.3%	

The odds ratio of 0.86 shows that females are about 14% less likely to drink lemonade than males, rather than 10% more likely as was shown earlier! This is an example of *Simpson's paradox*.

In general, Simpson's paradox illustrates that the effect of an omission of a categorical explanatory variable Z can have on the measure of association between a categorical explanatory variable X and a categorical response variable Y.

In the example, given the location variable Z, the conditional odds ratios show that the gender variable X and choice of drinks response variable Y have a positive association, with positive log-odds ratios. However, when the location variable Z is removed, the marginal association between X and Y is negative, with a negative log-odds ratio.

One reason for this apparent paradox is due to the dissimilar populations between the city and the rural groups. In the rural area, the majority of the test subjects are female, whereas in the city area, the majority is male.

For an excellent explanation of Simpson's paradox, please refer to the book below.

References

[1] A. Agresti, *An Introduction to Categorical Data Analysis*, Wiley & Sons, New York (1996).