Intelligent Data Analysis - Homework 2.2

Panagiotis Michalopoulos, Javier de la Rua, Michail Gongolidis,
Ignacio Rodriguez, Daniel Minguez

December 21st, 2018

1 Question 1

1.1 Obtain the estimated odds ratio and confidence intervals of crossing for car vs. truck at each traffic location.

We have calculated the odds ratio for car vs. truck depending on traffic, calculating the partial tables and applying the *oddsratio* function:

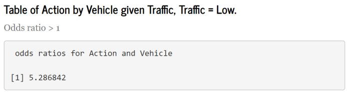


Table of Action by Vehicle given Traffic, Traffic = Low.

Odds ratio > 1

```
 odds ratios for Action and Vehicle

[1] 5.286842
```

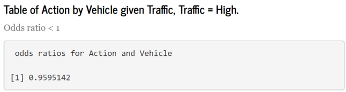


Table of Action by Vehicle given Traffic, Traffic = High.

Odds ratio < 1

```
 odds ratios for Action and Vehicle

[1] 0.9595142
```

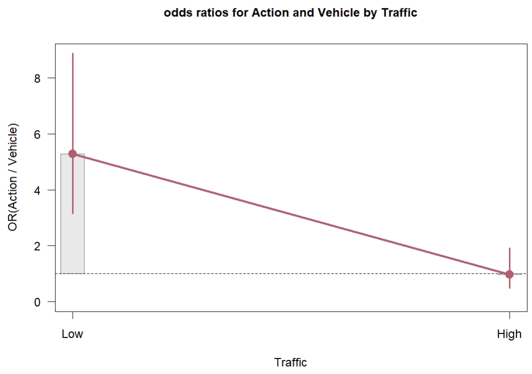
1.2 Interpret them and use the fourfold display to help you understand the output.

The odds ratio is greater than 1 for low traffic which means that the odds of an elk crossing when a car is passing are higher than the odds of crossing when it is a truck.

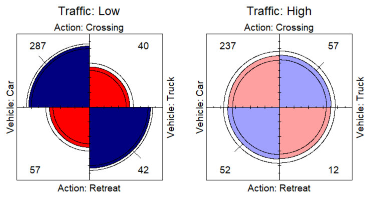
In case of high traffic the odds ratio is less than 1 which means that the odds of an elk crossing whit cars are lower than the odds for trucks.

Hoewer, since the confidence interval contains the value 1, we are going to assume that the odds ratio of crossing for cars is comparable to the odds ratio of crossing for trucks (with high traffic):

```
         2.5 %   97.5 %
Low  3.1495667 8.874459
High 0.4807475 1.915075
```



If we use the fourfold representation, we can see a graphical representation for the odds ratio of each traffic level. It includes confidence rings. If they don't overlap, that indicates an association between crossing and cars, every odds ration is significantly different from 0.

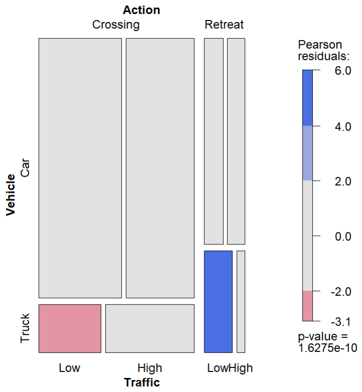


In the display we can see a strong positive association for low traffic (5.286842). The rings don't overlap and the non-principal diagonal sectors have less area than the principal diagonal ones (odds ratio greater than 1). On the other hand the odds of crossing is essentially identical for cars and trucks in case of high traffic (odds ration almost 1).

This result is an example of Simpson's paradox.

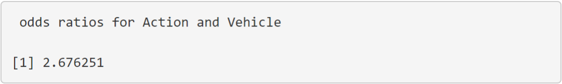
1.3 Use also the mosaic function to interpret partial tables.

From the mosaic plot below we see that there is no systematic association between different actions and the type of vehicle - except among the action of trucks under low traffic. The tiles show that there are relatively less trucks in low traffic with crossing that the hypothesis of independence would predict.



2 Obtain the estimated odds ratio of crossing vs. retreat without taking into account the third (control) variable. Would it be correct to exclude the effect of that third variable?

The odds of "success" (crossing) for car vehicles are higher than for trucks.



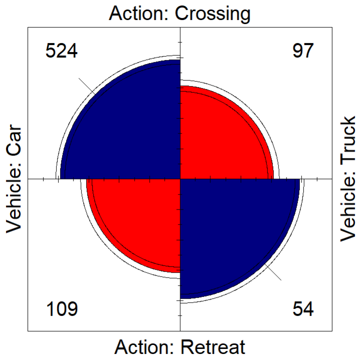
```
 odds ratios for Action and Vehicle

[1] 2.676251
```

Taking into account (controlling for) traffic, the odds of crossing are lower for car vehicles than for trucks in case of high traffic. Just the reverse direction that the marginal table showed.

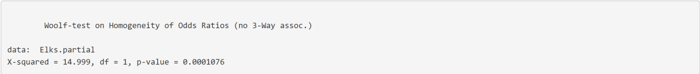
It´s an example of Simpson's Paradox: The result that a marginal association can have a different direction from each conditional association. It can be dangerous to "collapse" contingency tables over a third control variable.

In the Elks marginal table, where the odds ratio was 2.676251, we observe a strong positive association (odds ratio greater than 1).



3 Test the homogeneous association between X and Y controlling for Z (function woolf_test).

We have applied Woolf test on homogeneity of odds ratio:



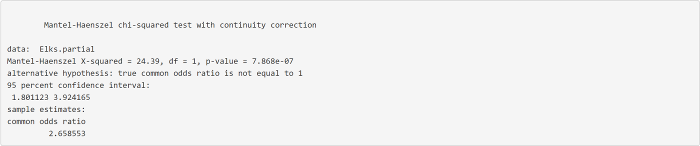
```
	Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)

data:  Elks.partial
X-squared = 14.999, df = 1, p-value = 0.0001076
```

The very low p-value (0.0001) indicates that we can reject the Homogeneity of the Odds Ratio through the two levels of traffic. We cannot assume that the conditional relationship between any pair of variables given the third one is the same at each level of the third variable.

4 Are X and Y conditional independent given Z?

We have used the Mantel-Haenszel chi-squared test of conditional indeendance:



```
	Mantel-Haenszel chi-squared test with continuity correction

data:  Elks.partial
Mantel-Haenszel X-squared = 24.39, df = 1, p-value = 7.868e-07
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.801123 3.924165
sample estimates:
common odds ratio
         2.658553
```

The very low p-value (7.868e-07) indicates that we can reject the conditional independence of the Odds Ratio through the two levels of traffic.