HW#9 – BZAN 535 – Categorical Data – Due Nov. 12, 2016

1. Using the dh journey data and a query similar to that on slide 5 of Lecture 22 (already posted on Bb):
    a. At the beginning of Lecture 21, we compute lift = 1.283 for bananas and soft drinks. Now compute the lift for this pair of products separately for sales < $10, $10-<$20, $20-<$40, $40-<$80, and ≥$80.
    b. Interpret the results for your calculations for lift at each basket size vs. the collective lift = 1.283. Explain why the difference arises. Which lift calculation best captures the affinity or lack thereof of bananas and soft drinks?

2. The buytest data summarize results for a test mailing of 10,000 catalogs. Purchase indicates whether the catalog recipient made a purchase or not.
    a. Using the buytest data, conduct separate chi-square tests for Purchase vs. Gender and Purchase vs. Married. In each case, state the conditional percentage of making a purchase for each level of the independent variable and comment on what the chi-square test leads one to conclude. (Note: 0 = No; 1 = Yes.)
    b. Compute the odds ratio for making a purchase depending on whether the catalog recipient is married or not; do this controlling for whether the recipient owns a home or not. That is, you will have two odds ratios, one for those who own a home (1) and another for those who are do not (0). Interpret each odds ratio carefully. Give an interval estimate for each true ratio. Is the association stronger for renters or for home owners? Explain.

3. For the dataset Household Spend, which summarizes the 801 Kroger shoppers for which we have demographic information, construct a contingency table for Income vs. Age. Since there are 12 income categories, and one has only 5 households, the initial table of counts will be too sparse. To remedy this problem, combine the top six categories into a single category labeled 100K+.

    Is there compelling evidence that, among (frequent) Kroger customers, income and age are not independent? Interpret the mosaic plot and summarize the relationship between these two variables. Discuss the cell(s) that most departs from what is expected under independence? (Note: Be sure to order the categories using the column property "Label ordering" so that the plot is useful, with age and income categories in correct order.) You can use JMP's "cell Chi-square" to spot cells with the most significant departure from independence.

4. Simpson's paradox – Lecture 22 ends with an introduction to the Bendix case study. Summarize in a paragraph the two different ways of looking at the data. Explain for a jury what is the reason for the seeming contradiction.