Homework 9 - Categorical Data

Kevin Gardner

Due Nov. 12, 2016

## Question 1

Using the dh journey data and a query similar to that on slide 5 of Lecture 22 (already posted on Bb):

1. 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.

****

Bananas and soft drinks

Basket size < $10



Basket size >= $10 and < 20



Basket size >= $20 and < 40



Basket size >= $40 and < 80



Basket size >= $80



1. 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?

As size of basket increases from less than $10 to $80+, the lift increases from about 0.5 to 1. For baskets of size $80 and up, we observe a lift of a lift of about 1.

For baskets under $10, the probability of having bananas is only half as big for baskets with softdrinks than it is for all baskets under $10.

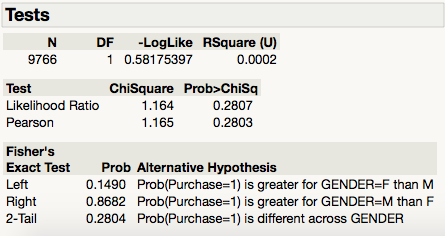
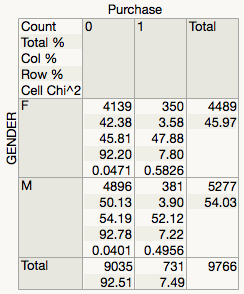
For baskets of $80 or more, the probability of having bananas is the same for baskets with softdrinks as it is for all baskets of that size.

For larger baskets there is evidence of a positive association. However, for smaller baskets, it is less likely that both items will be together in a basket than if there were no association. Controlling for basket size gives the truer picture of the association between bananas and soft drinks.

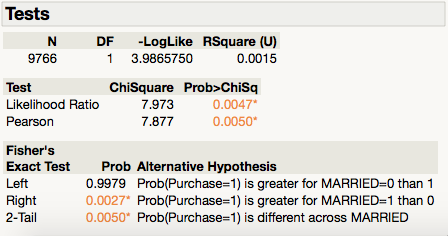
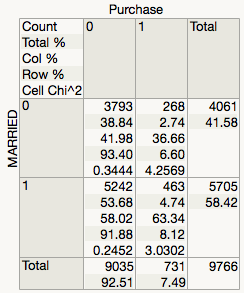
## Question 2

The buytest data summarize results for a test mailing of 10,000 catalogs. Purchase indicates whether the catalog recipient made a purchase or not.

1. 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.)



For females, 7.8% made a purchase; for males, only 7.22% made a purchase. The chi-square is 0.5826 for females and 0.4956 for males. The Pearson chi-square=1.165 (p=0.28), indicates that this difference is not statistically significant. Thus we are not convinced that gender makes any difference in the likelihood of purchase.



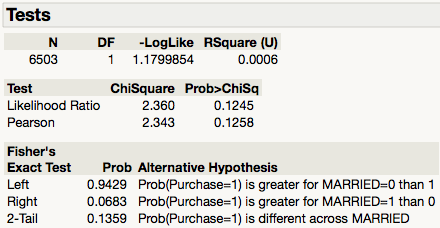
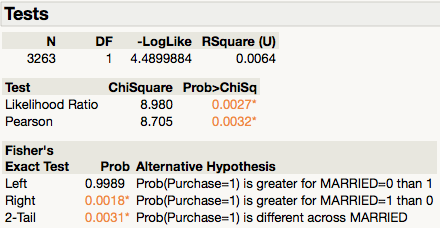
For non-married, 6.60% made a purchase; for married 8.12% made a purchase. The chi-square is 4.569 for non-married and 3.0302 for married.

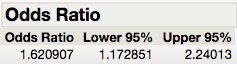
The Pearson chi-square=7.877 (p=0.005), indicates that a difference this great is highly unlikely unless there is a true difference in response rates due to marital status. Married recipients are more likely to make a purchase.

1. 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.

**Analysis of Purchase**

**By MARRIED OWNHOME=0 By MARRIED OWNHOME=1**

**Odds ratio**

Among catalog recipients who do not own a home, the odds of purchase are 1.15 times greater for individuals who are married versus not married; and we are 95% confident that the odds ratio is between 0.961437 and 1.375788.

Among catalog recipients who do own a home, the odds of purchase are 1.62 times greater for individuals who are married versus not married; and we are 95% confident the odds ratio is between 1.172851 and 2.24013.

**Chi-square**

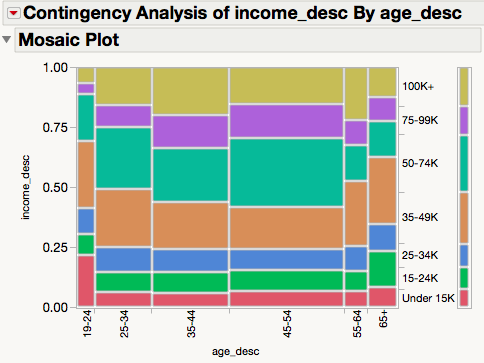
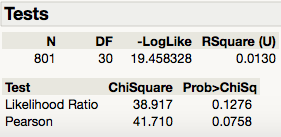
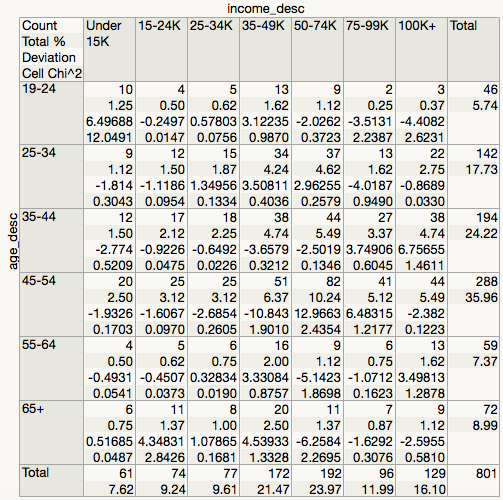
For recipients who do not own a home, the Likelihood Ratio and Pearson tests are both insignificant with chi-square of 2.360 and 2.343 respectively. So for this subset of catalog recipients the probability of purchase does not seem to depend marital status.

For recipients who do own a home, the Likelihood Ratio and Pearson tests are both significant with chi-square of 8.980 and 8.705 respectively. So for this subset of catalog recipients the probability of purchase does seem to depend marital status.

This suggests the association between purchase and marital status is stronger for home owners than it is for renters.

## Question 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.

The Likelihood Ratio and Pearson tests are both insignificant with chi-square of 0.1276 and 0.0758 respectively. We fail to reject the null hypothesis, so there is not compelling evidence that among (frequent) Kroger customers, income and age are not independent?

The largest deviation is age 45-54 with income 35K-49K (deviation=-10.843, chi-square=1.9010) and age 45-54 with income 50K-74K (deviation=12.9663, chi-square=2.4354).

From the mosaic plot we see the largest number of high frequency customers are 35-44 or 45-54 and have income 34K-49K or 50K-74K. In general, the number of high frequency customers increases and then decreases with both age and income.

## Question 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.

From plaintiff’s table we see that only 1 of the 14 employees in the “under 40” category got let go and 7 of the 31 employees in the 40 and over category were let go.

From the defendant’s table, which controls for job role, we see there were 0 staff engineers and 0 Sr. staff engineers in the “under 40” category and so the employees let go were all in the 40 and over category. Only 1 of the 14 project engineers in the “under 40” category got let go but 0 of the 5 project engineers in the 40 and over category were let go.

Thus we have Simpson’s paradox. The association between being let go and age category disappears when we control for job category.