Math 561 Assignment 4*

Due date: 13 Mar 2023.

Due date note

I know it's spring break on 3/13, however, I put that due date so you can have a few extra days to finish the homework instad of mkaing it due the Friday before, on 3/10. Hope this helps.

1. Solve exercise 9.5. from the book

Consider the data set displayed in the table below. This table appears in [Fin10] which is extracted from the original ecological study in [TCHN07]. The data set was used to search for a link between trophic levels and vegetable composition in rivers in northeast France. A river can be either oligotrophic, mesotrophic, or eutrophic if nutrient levels are poor, intermediate, or high, respectively. To each river, a triple of binary charac- ters was associated (r, p, e) indicating the presence or absence (1/0) of rare (r), exotic (e), or pollution-tolerant (p) species of vegetables. Two empty columns have been removed to leave a 3×6 table. Is there sufficient data to reject the null hypothesis that trophic level and vegetable composition of the rivers are independent?

Trophic level	(0,0,0)	(1,0,0)	(0,1,0)	(0,0,1)	(1,1,0)	(0,1,1)
Oligotrophic	0	0	3	0	3	2
Mesotrophic	2	1	0	2	1	0
Eutrophic	2	0	3	1	1	0

As you work on this problem, make sure you include all the details for the construction of the test of model fit:

- The null hypothesis is written for you; state clearly the log-linear model in question.
- What statistic are you using for goodness of fit?
- What method are you using to compute the distribution of this statistic?
- Write your conclusion using words. How do you think the data was generated, if the model doesn't fit this data? What does this mean in practice?

2. Zeros in tables

It is very important to understand the following two concepts of bounds on contingency table entries:

1. a sampling zero;
2. a structural zero.

A sampling zero means nobody was observed in the data in a particular cell of the table (in class example, we had an entire second column of a 2×3 table being zero). However, it's a property of the data collected, and does not necessarily mean there *cannot* be nonzero entries in those cells. A structural zero means that, in fact, there cannot be a nonzero entry. An easy example is if the table is classifying rates of various types of cancer between different genders, then men can't have ovarian cancer since they don't have ovaries - that's a structural constraint, not a data collection accident.

*Algebraic and Geometric Methods in Statistics, Spring 2023

In this problem you are to explore how structural and sampling zeros affect the goodness of fit analysis. Suppose we consider an example of the same structure:

	M	F	T/Nb	totals
$\leq$ 135K	?	?	?	**13**
$>$ 135K	?	?	?	**13**
totals	**10**	**10**	**6**	**26**

but we sample people from a company *whose policy is that men cannot be paid less than \$135K*. This is a structural zero on the table, do you agree?

- Pick a data table u with sample size 26 which fits these margins but also respects the structural zero.
- What is a Markov basis for the model of independence of 2×3? (Yes we wrote it down in class. Write it down again.)
- Can you apply all of the moves from this basis to your table u?
- Do you still think you can connect the fiber $\mathcal{F}(u)$ with the basis you had for the model? If yes, demonstrate that the zero doesn't present a problem. If no, demonstrate an example where Markov moves might "get stuck" or can't reach a particular table.
- What do you conclude?

Additional references for problem 2 Here is a full online lesson on the two types of zeros. There are many applications where zeros are very important; for example take a quick look at this applied paper. If you are curious about this, you can try to run the loglin method in `R` with the option of `start` which specifies the zeros.

Check out theorem 9.4.5 in the book (don't have to study it, just use it as a reference if you'd like to learn more).