Homework 2: Categorical Data

Nicholas G Reich, for Biostats 743 at UMass-Amherst

Your assignment should be submitted in two separate files by 5pm on Friday October 5th. The first, should be an RMarkdown (.Rmd) or another format that dynamically compiles your write up and runs the code inside it. The second should be the PDF file that was reproducibly compiled using the first file. All figures should be generated by the code, none should be loaded directly. The homework files should be submitted using your shared Google Drive folder with the instructor.

Inference for tables with small counts

One limitation of chi-squared tests that arises frequently in practice is that their inference is based on assumptions that may only be valid with large-sample sizes. A classical alternative to a chi-squared test in situations where there is a small cell count (i.e. less than 5) is to use Fisher's exact test. Another alternative would be to use Bayesian inference, which does not rely on large-sample approximations.

Question 1

Based on Table 1 from Denes et al (*American Journal of Epidemiology*, 1977, 105: 2), the authors were attempting to ascertain whether food-service workers were more likely to be sick when they had worked on a particular night where there had been a known outbreak of foodborne illness. [Acknowledgments to Brian Caffo for parts of this question.]

The data was

	Sick	Not sick
Worked	10	12
Did not work	2	26

(a) Obtain a copy of the original paper and read it. (Note: this is a test of your library science skills. Do not obtain a copy of this paper from a classmate. Locate this paper online either via typical search engine or using the UMass Libraries.)
(b) Re-run the chi-squared test that the authors ran in the paper to test the null hypothesis that employees who worked that night were as likely as those who did not to get sick.
(c) Run Fisher's exact test on this data to test the same hypothesis.
(d) Use a Bayesian analysis to test the same hypothesis.
(e) Write a paragraph comparing the assumptions and results of the three methods.

Question 2

Generalize the above setting to one where you have observations on $n = 50$ individuals and the contingency table has multinomial structure as follows:

	Sick	Not sick
$X = 1$	π_{11}	π_{12}
$X = 2$	π_{21}	π_{22}

with $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = (0.3, 0.3, 0.05, 0.35)$.

Conduct a simulation study comparing the performance of (1) a chi-squared test of independence testing whether the rows follow the same distribution, (2) Fisher's exact test, and the Bayesian method described in class using (3) uniform priors and (4) Jeffrey's priors for $\pi_{sick|x=2}$ and $\pi_{sick|x=2}$. Assuming a Type I error rate of 0.05, compare the statistical power of these four methods (i.e. how often does each method correctly reject the null hypothesis that $\pi_{sick|x=1} = \pi_{sick|x=1}$) using a simulation study. Assess the coverage rates for 80% and 95% posterior credible intervals from the Bayesian methods for the estimands (1) the relative risk of being sick comparing $X = 1$ to $X = 2$ and (2) the difference in conditional probabilities of being sick $\pi_{sick|x=2} - \pi_{sick|x=1}$. Summarize your findings in 1-2 paragraphs and 1-2 figures.

Contingency Table questions

Question 3

Test your computer's random normal generator. Simulate 10,000 random normal deviates and test whether or not they appear to be normal with a chi-squared test. Explain your steps and interpret your results. [Acknowledgments to Brian Caffo for this question.]

Question 4

The Chinese Mini-Mental Status Test (CMMS) is a test consisting of 114 items intended to identify people with Alzheimers disease (AD) and dementia among people in China. An extensive clinical evaluation was performed of this instrument, whereby participants were interviewed by psychiatrists and nurses and a definitive (clinical) diagnosis of AD was made. The table below shows the counts obtained on the sub-group of people with at least some formal education. Suppose a cutoff value of ≤ 20 on the test is used to identify people with AD. [Acknowledgments to Brian Caffo for this question.]

CMMS Score	No AD	AD
0-5	0	2
6-10	0	1
11-15	3	4
16-20	9	5
21-25	16	3
26-30	18	1

(a) What is the sensitivity and specificity of the CMMS test using the 20 cutoff?
(b) Graph the positive predictive value as a function of the prevalence of AD. Do the same for the negative predictive value.
(c) What would be the sensitivity and specificity if a cutoff of 15 had been used?