Homework 2: Categorical Data

Nicholas G Reich, for Biostats 743 at UMass-Amherst

Your assignment should be submitted in two separate files by 8am on Friday September 29th. The first, should be an RMarkdown (.Rmd) or LaTeX (.tex) file. The second should be the PDF file that was reproducibly compiled using the first file. 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 (1977), 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.

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.
(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_{21}
$X = 2$	π_{12}	π_{22}

with $(\pi_{11}, \pi_{21}, \pi_{21}, \pi_{22}) = (0.2, 0.4, 0.05, 0.35)$.

Conduct a simulation study comparing the performance of (1) a chi-squared test of independence, (2) Fisher's exact test, and (3) the simple Bayesian method described in class. Compare Type I error rates between the

three methods. 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.

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.]

```
- Bayesian example, BDA3 p 37
- chisq GOF test on draws from some distribution
- derivation of chi-sq being equivalent to square of test of proportions test stat
- 652 HW3, problem 6: likelihood interval comparison to
- Poisson GOF test
- Bayesian multinomial example
```