Homework 1: Categorical Data

Nicholas G Reich, for Biostats 743 at UMass-Amherst

Your assignment should be submitted in two separate files by 5pm on Friday September 21st. 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. Before submitting your files, create a folder on Google Drive with the name "[lastname]-[firstname]-cda" (e.g. "Reich-Nicholas-cda") and share it with the TA at `zhengfanwang@umass.edu`. The homework files should then be submitted by copying the files into this folder.

Question 1

Say you flip a coin 100 times and get 53 heads.

a. Write down the Score test statistic and by hand (i.e. don't used a canned R function to calculate the test statistic for you) conduct a test of the null hypothesis $H_0 : \pi$=0.5.
b. Draw a picture of the likelihood and show what features of the likelihood are used for the Score test statistic.
c. Explain (using words, pictures, and equations) how the Score test uses different information than the likelihood ratio test.

Question 2

Conduct a simulation study to evaluate the coverage probabilities of different methods for computing confidence intervals for π. You may vary the true π, the sample size, and must choose at least 4 methods to compare. Your results should be summarized with a one paragraph description of the implementation of the simulation study, a one paragraph descirption of the results, and one figure (multi-panel is ok) summarizing the results.

Question 3

Prove that if $Y \sim Binom(n, \pi)$ and $\pi \sim beta(\alpha_1, \alpha_2)$ (for $\alpha_1 > 0$ and $\alpha_2 > 0$) then the posterior $h(\pi|y)$ is a beta distribution, specifically, h follows a $beta(y + \alpha_1, n - y + \alpha_2)$.

Question 4

Complete problem 1.17 in CDA.

Suppose that $P(Y_i = 1) = 1 - P(Y_i = 0) = \pi, i = 1, .., n$, where $\{Y_i\}$ are independent. Let $Y_i = \sum_i Y_i$.

(a) What is the distribution of Y? What are $E(Y)$ and $var(Y)$?

(b) When $\{Y_i\}$ instead have pairwise correlation $\rho > 0$, show that $var(Y) > n\pi(1 - \pi)$, overdispersion relative to the binomial. Altham (1978) and Ochi and Prentice (1984) discussed generalizations of the binomial that allow correlated trials.

(c) Suppose that heterogeneity exists: $P(Y_i = 1|\pi) = \pi$ for all i, but π is a random variable with density function $g()$ on $[0, 1]$ having mean ρ and positive variance. Show that $var(Y) > n\rho(1 - \rho)$. (When π has a beta distribution, Y has the betabinomial distribution of section 14.3)

Question 5

Complete problem 1.3 in CDA.

An experiment studies the number of insects that survive a certain dose of an insecticide, using several batches of insects of size n each. The insects are sensitive to factors that vary among batches during the experiment but were not measured, such as temperature level. Explain why the distribution of the number of insects per batch surviving the experiment might show overdispersion relative to a $bin(n, \pi)$ distribution.

Question 6

Complete problem 1.22 in CDA.

Suppose that $y_1, y_2, ..., y_n$ are independent from a Poisson distribution.

(a) Obtain the likelihood function. Show that the ML estimator $\hat{\mu} = \bar{y}$.

(b) Construct a large-sample test statistic for $H_0 : \mu = \mu_0$ using (i) the Wald method, (ii) the score method, and (iii) the likelihood-ratio method.

(c) Explain how to construct a large-sample confidence interval for μ using (i) the Wald method, (ii) the score method, and (iii) the likelihood-ratio method.

Question 7

Complete problem 1.29 in CDA.

Genotypes AA, Aa, and aa occur with probabilities $[\theta^2, 2\theta(1-\theta), (1-\theta)^2]$. A multinomial sample of size n has frequencies (n_1, n_2, n_3) of these three genotypes. (a) Form the log likelihood. Show that $\hat{\theta} = (2n_1 + n_2)/(2n_1 + 2n_2 + 2n_3)$.

(b) Show that

$$\frac{\partial^2 L(\theta))}{\partial \theta^2} = [(2n_1 + n_2)/\theta^2] + [(n_2 + 2n_3)/(1-\theta)^2]$$

and that its expectation is $2n/\theta(1^{-}\theta)$. Use this to obtain an asymptotic standard error of $\hat{\theta}$.

(c) Explain how to test whether the probabilities truly have this pattern.

Question 8

Complete problem 1.30 in CDA.

Refer to Section 1.5.6 and the model for pneumonia infections in calves. Using the likelihood function to obtain the information, show that the approximate standard error of $\hat{\pi}$ is $\sqrt{\pi(1-\pi)/n(1+\pi)}$.