Biometry 711 - Categorical Data Analysis
Summer 2015
Homework 1
Due Friday 19 June 2015, 6pm

1. Consider a random sample of n independent Bernoulli(π) observations.

   (a) Show $\text{var}(\hat{\pi}) = \text{var}(y/n) = \pi(1-\pi)/n$, where y is the total number of successes in the n trials. On the same set of axes, graph $\text{var}(\hat{\pi})$ as a function of π using $n = 5$ and $n = 20$. For what value(s) of π is $\text{var}(\hat{\pi})$ greatest? least? Explain why this makes sense.

   (b) A common study objective is to estimate the probability of an event or trait in a population. The sample size justification for such an objective entails determining the precision with which the unknown probability can be estimated, where precision is measured by the half-width of the corresponding 95% confidence interval (commonly known as the *margin of error*). Typically, the Wald CI is used, primarily because of its ease of use (Equation 1.13). However, since the true event probability is unknown and, since we're only in the planning stages of the study and no data have yet been collected, neither π nor $\hat{\pi}$ are known. Nonetheless, an upper bound on the margin of error can still be estimated. Explain how your findings in question 1a allow you to calculate this upper bound. For a sample size of $n = 20$, what is the maximum margin of error based on the Wald interval?

   (c) Because the sample size is small, you might also plan the study based on the width of the corresponding exact binomial (Clopper-Pearson) 95% CI (Section 16.6.1). (Note that since the exact interval is not necessarily symmetric about $\hat{\pi}$, we report CI width rather than half-width.) Using the `R` function `binom.test` (or otherwise if you prefer `SAS`), calculate the maximum width for an exact 95% CI of a probability based on a sample size of $n = 20$? How does this compare to the results of question 1b?

2. Exercise 2.4

3. Exercise 2.17

4. Exercise 2.37

5. Let C_i be the number of incident cases of a rare disease among n_i individuals at risk in geographic region i, $i = 1, \ldots, I$, with observed value c_i. Consider a null hypothesis that the cases do not 'cluster,' where 'clustering' means the spatial aggregation of cases occurs non-randomly. Then the null hypothesis can be stated as $H_0 : \{C_i\}$ are independent Poisson random variables with $E(C_i) = \lambda n_i$, where λ is the baseline incidence rate. Therefore, under the null hypothesis of no clustering, the expected number of cases in geographic location i is the baseline incidence rate of disease multiplied by the number at risk. If the baseline rate, λ, is unknown, a conditional null hypothesis can be constructed based on the distribution of $\{C_i\}$ conditional on the sufficient statistic $c_+ = \sum_i c_i$. Show the distribution of $\{C_i\}|c_+$ does not depend on λ. What is this conditional distribution?

6. Consider the relative risk, $\rho = \pi_1/\pi_2$ (Equation 2.3), with estimate $r = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{y_1/n_1}{y_2/n_2}$. As stated on page 71, $\log(r)$ converges faster to normality than r.

(a) Find the asymptotic distribution of $\log(r)$. Show all work.

(b) Derive the Wald test for the null hypothesis $H_0 : \log(\rho) = 0$. State your final answer as a z-test. Show all work.

(c) Invert the Wald test to derive the form of a $(1 - \alpha)100\%$ Wald confidence interval for $\log(\rho)$. Show all work.

(d) Using the result stated in question 6c, derive a $(1 - \alpha)100\%$ confidence interval for ρ.

7. Consider the null hypothesis $H_0 : \pi_1 = \pi_2$ and $\pi_3 = \pi_4$, where $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)'$ is the parameter vector for the four-class multinomial distribution. Derive the Score test for this null hypothesis.