Categorical Data Analysis: HW 2

1. Using R, compute the coverage probability of the score, Clopper-Pearson and Bayesian highest posterior (with Jeffreys prior, i.e., $Beta(1/2, 1/2)$) interval for the binomial proportion π when $n = 20$. To compute the coverage probability, find the confidence interval for each of the possible $n+1$ outcomes of the binomial trial. Then, see which of the $n+1$ C.I.'s contain π. To find the coverage probability, you then simply add up the binomial probabilities for those intervals that contain π. In other words, the coverage probability for a confidence interval is given by
$$\sum_{\text{y: C.I.contains } \pi} \binom{n}{x} \pi^y (1-\pi)^{n-y},$$
where the sum is over all those CI for the given y that contain π.

   Compute the coverage probability for at least 4 different values of π and for all three methods of finding an interval mentioned above. If you want extra credit, you turn in a nice graph that has π on the x-axis (use a lot of different values so that you get a nice plot) and the coverage probability for each of the three methods (in a different color, but all in the same plot) on the y-axis. For even more extra credit, show the same plot but now for $n = 10$ and $n = 50$.

2. (Agresti, chapter 1) A binomial experiment tests $H_0 : \pi = 0.50$ against $H_A : \pi \neq 0.50$ using significance level $\alpha = 0.05$. Only n = 5 observations are available. Show that the actual type I error for the exact binomial test is zero and the actual type I error for the score test is 1/16.

3. This exercise proofs the important fact that it really doesn't matter what sampling scheme one uses (Poisson, multinomial or product multinomial), one always ends up with the same likelihood function and hence inference. To tackle this problem, let's only apply it to 2×2 tables, with cell counts N_{11}, N_{12}, N_{21} and N_{22}.

   (a) Write down the joint distribution function of these 4 cell counts assuming Poisson sampling, i.e. $N_{ij} \sim Pois(\mu_{ij})$.

   (b) Find the distribution of $N = \sum_i \sum_j N_{ij}$.

   (c) Find the conditional distribution function of the cell counts N_{11}, N_{12}, N_{21} and N_{22} given that the total sample size $N = n$, i.e., find $P(N_{11} = n11, N_{12} = n12, N_{21} = n21, N_{22} = n22 | N = n)$.

   (d) Argue that this is exactly the likelihood for multinomial sampling with what formula for the π_{ij}'s?

4. (Agresti, chapter 2) For a diagnostic test of a certain disease, let π_1 denote the probability that the diagnosis is positive given that a subject has the disease (called the *sensitivity*), and let π_2 denote the probability that the diagnosis is positive given that a subject does not have it. ($1 - \pi_2$, the probability of the

test being negative given that the subject does not have the disease is called the *specificity.*) Let ρ denote the probability that a subject has the disease.

(a) More relevant to a patient who has received a positive diagnosis is the probability that he or she truly has the disease. Given that a diagnosis is positive, show that the probability that a subject has the disease (called the positive predictive value) is

$$\pi_1\rho/[\pi_1\rho + \pi_2(1-\rho)]$$

(b) Suppose that a diagnostic test for HIV + status has both sensitivity and specificity equal to 0.95, and $\rho = 0.005$. Find the probability that a subject is truly HIV+, given that the diagnostic test is positive.

(c) To better understand the answer in (b) find the joint probabilities in the 2×2 table relating diagnosis to actual disease status and discuss their relative sizes.

(d) Discuss how the answer in (b) depends on the prevalence ρ. Illustrate by finding the answer when $\rho = 0.10$ instead of 0.005.