Psy 5011: Categorical Data Analysis **(Due: 10/03/20)**

Homework #1

Instructor: Yao, Kaiping Grace *Name:* Hao-Cheng Lo, *Id:* D08227124

Problem 1.4: In a particular city, the population proportion π supports an increase in the minimum wage. For a random sample of size 2, let Y = number who support an increase.

(a) Assuming $\pi = 0.50$, specify the probabilities for the possible values y for Y and find the distribution's mean and standard deviation.

The random variable Y obeys binominal distribution:

$$Y \sim B(N = 2, \pi = 0.5)$$

So, here are the probabilities for the possible values y for Y:

$$P(Y = y) = \binom{N}{y} \pi^y (1 - \pi)^{N-y}$$

$$P(Y = 0) = \binom{2}{0} 0.5^0 0.5^2 = 0.25$$

$$P(Y = 1) = \binom{2}{1} 0.5^1 0.5^1 = 0.5$$

$$P(Y = 2) = \binom{2}{2} 0.5^2 0.5^0 = 0.25$$

The mean and standard deviation are:

$$\mathrm{E}(Y) = N\pi = 2 \times 0.5 = 1$$

$$\mathrm{Var}(Y) = N\pi(1 - \pi) = 2 \times 0.5 \times 0.5 = 0.5$$

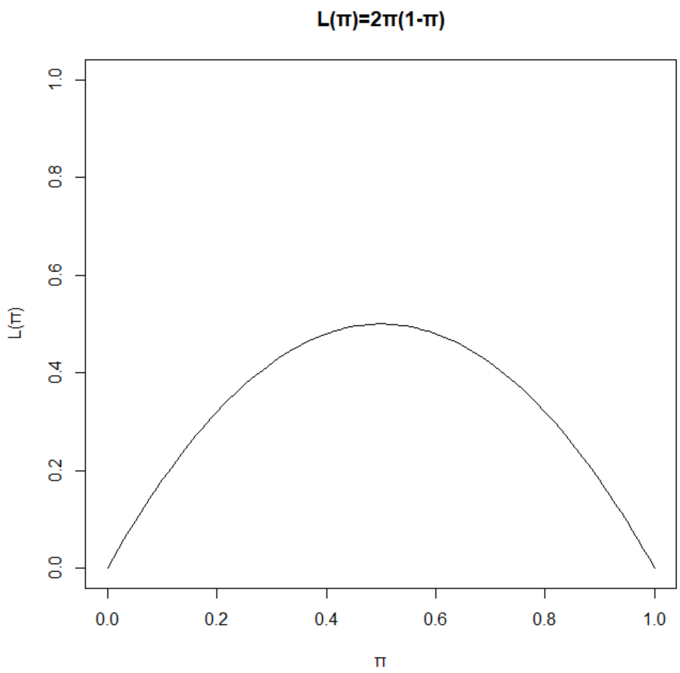
$$\mathrm{SD}(Y) = \sqrt{0.5} = .71$$

(b) Suppose you observe $y = 1$ and do not know π. Find and sketch the likelihood function. Using the plotted likelihood function, explain why the ML estimate $\hat{\pi} = 0.50$.

Given $y = 1$ and $N = 2$, the likelihood function of the binomial model is:

$$L(\pi|N = 2, y = 1) = \binom{2}{1} \pi(1 - \pi)$$

Let $\pi = \{0.00, 0.01, 0.02, 0.03, ..., 0.99, 1.00\}$ with total 100 points, plot the $\pi - L(\pi|N = 2, y = 1)$ relation with R:



By observing the plot or by letting $L'(\pi|N = 2, y = 1) = 2\pi + 1 = 0$, we can find that maximum likelihood occurs at $\pi = 0.5$ (equals to analytical solution of MLE of binominal distribution $\hat{\pi} = y/N$), so we can let $\hat{\pi} = 0.5$. Therefore, 0.5 is the most likely given the data is the MLE of π.

Problem 1.6: Genotypes AA, Aa, and aa occur with probabilities (π_1, π_2, π_3). For $n = 3$ independent observations, the observed frequencies are (y_1, y_2, y_3).

(a) Explain how you can determine y_3 from knowing y_1 and y_2. Thus, the multinomial distribution of (y_1, y_2, y_3) is actually two-dimensional.

Because total number of observed frequencies n is fixed (known), so we can depict (y_1, y_2, y_3) as $(y_1, y_2, n - y_1 - y_2)$. Because the parameters remain y_1 and y_2, thus the multinomial distribution of (y_1, y_2, y_3) is actually two-dimensional.

(b) Show the ten possible observations (y_1, y_2, y_3) with $n = 3$.

10 possible obs $= \{(0, 0, 3), (0, 1, 2), (0, 2, 1), (0, 3, 0), (1, 0, 2), (1, 1, 1), (1, 2, 0), (2, 0, 1), (2, 1, 0), (3, 0, 0)\}$

(c) Suppose $(\pi_1, \pi_2, \pi_3) = (0.25, 0.50, 0.25)$. What probability distribution does y_1 alone have?

For y_1 alone, we can reduce the multinominal distribution of $Y \sim multinominal(N = 3, \pi_1 = 0.25, \pi_2 = 0.5, \pi_3 = 0.25)$ to binominal distribution of $Y \sim B(N = 3, \pi_1 = 0.25)$. This is because for y_1 alone, any independent observation is a Bernoulli trial (y_1 is Yes or NO) where $\pi = 0.25$. Considering the $N = 3$, for y_1 alone, this is a binominal distribution of $N = 3,\ \pi_1 = 0.25$

Problem 1.9: A study of 100 women suffering from excessive menstrual bleeding considers whether a new analgesic provides greater relief than the standard analgesic. Of the women, 40 reported greater relief with the standard analgesic and 60 reported greater relief with the new one.

(a) Test the hypothesis that the probability of greater relief with the standard analgesic is the same as the probability of greater relief with the new analgesic. Report and interpret the P-value for the two-sided alternative. (Hint: Express the hypotheses in terms of a single parameter. A test to compare matched-pairs responses in terms of which is better is called a sign test.)

Let π be proportion of women suffering menstrual bleeding get relief with the new analgesic. By CLT, when N is large enough (> 20), a reasonable approximation to $B(N, \pi)$ is given by the normal distribution $N(\pi, \pi(1-\pi)/N)$. Given $\alpha = .05$.

$$H_0 : \pi = .5,\ H_a : \pi \neq .5$$

Test statistic:

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/N}} = \frac{.6 - .5}{\sqrt{.5(.5)/100}} = 2$$

The p-value for $z = 2.00$ is .022 is less than $\alpha/2 = .025$, so we reject H_0 which the probability of greater relief with the standard analgesic is not the same as the probability of greater relief with the new analgesic.

(b) Construct and interpret a 95% confidence interval for the probability of greater relief with the new analgesic.

The 95% CI of the probability of greater relief with the new analgesic (following (a), using null standard error):

$$p \pm z_{\alpha/2}(SE) = .6 \pm 1.96\sqrt{.5(.5)/100} \rightarrow (.502, .698)$$

A 95% confidence interval $(.502, .698)$ is a range of values that we can be 95% certain contains the true π. That is, if the current study repeats this process many times, then about 95% of the intervals produced will capture the true proportion π of women suffering menstrual bleeding get relief with the new analgesic.

Problem 1.10: Refer to the previous exercise. The researchers wanted a sufficiently large sample to be able to estimate the probability of preferring the new analgesic to within 0.08, with confidence 0.95. If the true probability is 0.75, how large a sample is needed to achieve this accuracy? (Hint: For how large an n does a 95% confidence interval have margin of error equal to about 0.08?)

$$z_{\alpha/2}(SE) = 1.96\sqrt{.75(.75)/n} = 0.08$$

$$n = 337.6 \approx 338$$

Hence, we need 338 sample size to achieve the accuracy.