Math 2a Homework 7 Solutions

Problem 1. (*Moore & McCabe, 8.48*) Let p_1 denote the proportion of men with normal chromosomes that have a criminal record, and let p_2 denote the proportion of men with abnormal chromosomes that have a criminal record. Our hypotheses are

$$H_0 : p_1 = p_2$$
$$H_a : p_1 < p_2.$$

In the notation of Moore & McCabe we have $n_1 = 4096$ and $\hat{p}_1 = 381/4096 = 0.0930$, while we have $n_2 = 28$ and $\hat{p}_2 = 8/28 = 0.2857$. We set $D = \hat{p}_1 - \hat{p}_2 = -0.1927$ and compute

$$\hat{p} = \frac{381 + 8}{4124} = 0.0943$$

and

$$SE_{D_p} = \sqrt{0.0943 \times 0.9057 \times \left(\frac{1}{4096} + \frac{1}{28}\right)} = 0.0554.$$

Our z statistic is given by

$$z = \frac{-0.1927}{0.0567} = -3.4783.$$

Therefore our P value is

$$P(Z \leq -3.4783) = 0.0003.$$

Since our P value is so small we confidently reject H_0 in favor of H_a. That is, we conclude that men with abnormal chromosomes are more likely to have a criminal record. However, one cannot conclude from this that chromosome abnormalities are a direct cause of increased criminality. There may be many other factors associated with chromosome abnormalities, that are not discussed in this study, which cause an increase in criminality. For example people with chromosome abnormalities may receive less education or have less money than the rest of the population.

Problem 2. (a) Of the population of Jewish people dying within a week of Passover, let p denote the proportion dying before Passover. Our hypotheses are

$$H_0 : p = 1/2$$
$$H_a : p < 1/2.$$

For our sample of size 1919 we compute

$$\hat{p} = \frac{922}{1919} = 0.4805.$$

Our z statistic is given by

$$z = \frac{0.4805 - 0.5}{\sqrt{\frac{0.5^2}{1919}}} = -1.7084.$$

Our P value is therefore

$$P(Z \leq -1.7084) = 0.044.$$

Thus we reject the null hypothesis at all significance levels ≥ 0.044. This sample gives evidence toward rejecting the null hypothesis.

(b) We repeat the hypothesis test from part (a) with the population of men with Chinese and Japanese ancestry who die within a week of Passover. We take p to be the proportion of such people dying before Passover. Again our hypotheses are

$$H_0 : p = 1/2$$
$$H_a : p < 1/2.$$

For our sample of size 852 we compute

$$\hat{p} = \frac{418}{852} = 0.4906.$$

Our z statistic is given by

$$z = \frac{0.4906 - 0.5}{\sqrt{\frac{0.5^2}{852}}} = -0.5488.$$

Our P value is therefore

$$P(Z \leq -0.5488) = 0.29.$$

With such a high P value we do not have sufficient evidence to reject the null hypothesis.

(c) Presumably there are not too many men of Chinese or Japanese ancestry who are Jewish. Thus, we have no reason to expect the alternative hypothesis to hold in part (b) of this problem. In particular if we had computed a small P value in part (b) then this would lead us to suspect that there may be other, non-Jewish, factors that cause more people to die the week after Passover than the week after.

Problem 3. Let X have a binomial distribution with n trials and probability p of success. In order to compute the maximum likelihood estimate for p we need to find the value of p which maximizes the probability $P(X = x \mid p)$ for each possible outcome x. Since the distribution is binomial we have

$$P(X = x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

It is clear that if $x = 0$ this function is maximized when $p = 0$ and if $x = n$ then it is maximized when $p = 1$. Now suppose that $1 \leq x \leq n-1$. With x fixed, it is clear that p maximizes the function $P(X = x \mid p)$ if and only if maximizes the function

$$g(p) = p^x (1-p)^{n-x}.$$

Differentiating g with respect to p gives

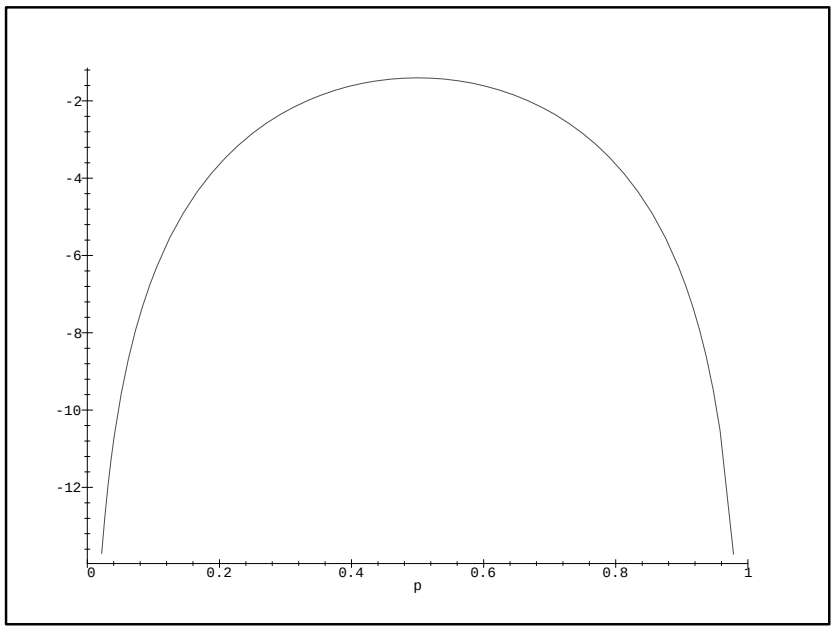
$$\begin{aligned} g'(p) &= xp^{x-1}(1-p)^{n-x} - (n-x)p^x(1-p)^{n-x-1} \\ &= p^{x-1}(1-p)^{n-x-1}(x - np). \end{aligned}$$

Thus if $g'(p) = 0$ then $p = 0$, 1 or x/n. Clearly the function $g(p)$ is not maximized when $p = 0$ or 1, since in both these cases $g(p) = 0$ (recall we are assuming that $x \notin \{0, n\}$). Since our function g is continuous and non-negative on $[0, 1]$ we deduce that the maximum occurs when $p = x/n$. Thus we see that for all values of x, the maximum likelihood estimate for p is x/n. Hence as a function of the random variable X, the maximum likelihood estimate for p is X/n.

Suppose now that $n = 10$ and $X = 5$. The log-likelihood function, $\ell(p)$, is given by the formula

$$\ell(p) = \log(P(X = 5 \mid p)) = \log\left(\binom{10}{5} p^5 (1-p)^5\right).$$

This function is plotted below.



Problem 4. Let N denote the total number of animals. Of these N, 100 are tagged. Suppose we recapture 50 animals, we let X denote the random variable that gives the number of these 50 animals which are tagged. We will assume that the 100 animals chosen to be tagged were chosen uniformly at random from the population, and that the second sample of 50 animals was also chosen uniformly at random from the population. Under these assumptions X has a hypergeometric distribution (cf Example I on page 13 of Rice). Thus the probability of $X = 20$ is given by the formula

$$\frac{\binom{100}{20}\binom{N-100}{30}}{\binom{N}{50}}.$$

Viewed as a function of N this is the likelihood $\mathrm{lik}(N)$. We note that from the data we have we necessarily have $N \geq 130$. We wish to find the value of N which maximizes this function. We have

$$\frac{\mathrm{lik}(N)}{\mathrm{lik}(N-1)} = \frac{(N-100)(N-50)}{N(N-130)}.$$

Now $\mathrm{lik}(N) > \mathrm{lik}(N-1)$ if and only if this ratio is greater than 1, which is if and only if

$$\begin{aligned}(N-100)(N-50) &> N(N-130)\\ N^2 - 150N + 5000 &> N^2 - 130N\\ 5000 &> 20N.\end{aligned}$$

Therefore we see that $\mathrm{lik}(N) > \mathrm{lik}(N-1)$ if and only if $N < 250$. In the same way we find that $\mathrm{lik}(N) < \mathrm{lik}(N-1)$ if and only if $N > 250$. Hence $\mathrm{lik}(N)$ is maximized when $N = 249$ or 250. Thus we estimate the population to be 250 (or 249).

Problem 5. (a) Let X_1 be the random variable that counts the number of starchy and green plants, X_2 the random variable that counts the number of starchy and white plants, X_3 and X_4 are defined similarly. Then the function we wish to maximize is

$$f(1997, 906, 904, 32 \mid \theta)$$

where $f(X_1, X_2, X_3, X_4 \mid \theta)$ is the joint distribution of the X_i. This distribution is multinomial and we have

$$f(1997, 906, 904, 32 \mid \theta) = \frac{3839!}{1997!906!904!32!}\left(\frac{2+\theta}{4}\right)^{1997}\left(\frac{1-\theta}{4}\right)^{906}\left(\frac{1-\theta}{4}\right)^{904}\left(\frac{\theta}{4}\right)^{32}.$$

Clearly, maximizing this function is equivalent to maximizing the function

$$g(\theta) = (2+\theta)^{1997}(1-\theta)^{906}(1-\theta)^{904}\theta^{32}.$$

Taking logs this is equivalent to maximizing the function

$$h(\theta) = \log g(\theta) = 1997\log(2+\theta) + 906\log(1-\theta) + 904\log(1-\theta) + 32\log\theta.$$

Differentiating h with respect to θ and multiplying by $(2+\theta)(1-\theta)\theta$ we find that the derivative of h vanishes precisely if

$$1997(1-\theta)\theta - 906(2+\theta)\theta - 904(2+\theta)\theta + 32(2+\theta)(1-\theta) = 0$$

which is if

$$-3839\theta^2 - 1655\theta + 64 = 0.$$

The roots of this polynomial are $\theta = -0.4668$ and $\theta = 0.0357$. Since we are given $0 < \theta < 1$ we see that the maximum likelihood estimate for θ is given by $\hat{\theta} = 0.0357$.

(b) In order to construct such an interval we need to compute $I(\hat{\theta}) = I(0.0357)$. We recall that

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1, X_2, X_3, X_4 \mid \theta)\right].$$

Now we have

$$f(X_1, X_2, X_3, X_4 \mid \theta) = \frac{3839!}{X_1!X_2!X_3!X_4!}\left(\frac{2+\theta}{4}\right)^{X_1}\left(\frac{1-\theta}{4}\right)^{X_2}\left(\frac{1-\theta}{4}\right)^{X_3}\left(\frac{\theta}{4}\right)^{X_4}$$

and therefore $\log f(X_1, X_2, X_3, X_4 \mid \theta)$ is given by

$$\log 3839! - \sum_{i=1}^{4}(\log X_i! - X_i\log 4) + \log(2+\theta)X_1 + \log(1-\theta)X_2 + \log(1-\theta)X_3 + \log(\theta)X_4.$$

Thus we have

$$\frac{\partial^2}{\partial\theta^2}\log f(X_1, X_2, X_3, X_4 \mid \theta) = -\frac{1}{(2+\theta)^2}X_1 - \frac{1}{(1-\theta)^2}X_2 - \frac{1}{(1-\theta)^2}X_3 - \frac{1}{\theta^2}X_4.$$

Each X_i is a binomial distribution with the appropriate parameters we compute

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1, X_2, X_3, X_4 \mid \theta)\right] = \frac{3839}{4}\left(\frac{1}{(2+\theta)} + \frac{1}{(1-\theta)} + \frac{1}{(1-\theta)} + \frac{1}{\theta}\right).$$

The estimated standard error of $\hat{\theta}$ is given by

$$s_{\hat{\theta}} = \frac{1}{\sqrt{I(\hat{\theta})}} = \frac{1}{\sqrt{25364.65}} = 0.0058.$$

An approximate 95% confidence interval for θ is given by $\hat{\theta} \pm 1.96 s_{\hat{\theta}}$, or $(0.024, 0.047)$.

(c) We have

$$E\left(\frac{4X_1}{n} - 2\right) = \frac{4}{n}E(X_1) - 2 = \theta$$

and

$$E\left(\frac{4X_4}{n}\right) = \frac{4}{n}E(X_4) = \theta.$$

Hence these are both unbiased estimates of θ. We have

$$\text{Var}\left(\frac{4X_1}{n} - 2\right) = \frac{16}{n^2}\text{Var}(X_1) = \frac{4-\theta^2}{n}$$

and

$$\text{Var}\left(\frac{4X_4}{n}\right) = \frac{16}{n^2}\text{Var}(X_4) = \frac{(4-\theta)\theta}{n}.$$

The asymptotic variance of the maximum likelihood estimate is given by

$$-\frac{1}{E[\ell''(\theta)]}$$

which as computed above is given by the formula

$$\left[\frac{n}{4}\left(\frac{1}{(2+\theta)} + \frac{1}{(1-\theta)} + \frac{1}{(1-\theta)} + \frac{1}{\theta}\right)\right]^{-1}.$$

This simplifies to give the asymptotic variance of the maximum likelihood estimate as

$$\frac{2\theta(1-\theta)(2+\theta)}{(1+2\theta)n}.$$

We note that the asymptotic variance of the maximum likelihood estimate for θ is less than the two variances computed above for all possible values of θ.