James K. Pringle
140.674 Stat Theory
Dr. Constantine Frangakis
Assignment 6
May 28, 2013

Assignment 6

The microarray experiment

1. Let Z be a Bernoulli random variable with probability $\mathrm{pr}(Z = 1) = \pi$; let Y be such that, if $Z = 1$, then Y is a draw from a standard normal distribution; but if $Z = 0$, then Y is a draw from a normal distribution with mean μ and variance 1.

   (i) Find $E(Y \mid \mu, \pi)$.

   Proof. The problem defines Y as a mixture distribution. We have

   $$Y = \pi N(0,1) + (1-\pi)N(\mu,1) \tag{1}$$

   By the linearity of expected value, we have

   $$\begin{aligned} E(Y \mid \mu,\pi) &= E(\pi N(0,1) + (1-\pi)N(\mu,1)) && (2)\\ &= \pi E(N(0,1)) + (1-\pi)E(N(\mu,1)) && (3)\\ &= (1-\pi)\mu && (4) \end{aligned}$$

   □

   (ii) Find $E(Y^2 \mid \mu, \pi)$.

   Proof. Let φ_{μ,σ^2} denote the density of $N(\mu,\sigma^2)$. By Wikipedia (mixture distribution), the density of Y as in (1) is

   $$f_Y(y) = \pi\varphi_{0,1}(y) + (1-\pi)\varphi_{\mu,1}(y) \tag{5}$$

   Therefore,

   $$\begin{aligned} E(Y^2 \mid \mu,\pi) &= \int_{-\infty}^{\infty} y^2 f_Y(y)dy && (6)\\ &= \int_{-\infty}^{\infty} y^2(\pi\varphi_{0,1}(y) + (1-\pi)\varphi_{\mu,1}(y))dy && (7)\\ &= \pi\int_{-\infty}^{\infty} y^2\varphi_{0,1}(y)dy + (1-\pi)\int_{-\infty}^{\infty} y^2\varphi_{\mu,1}(y)dy && (8)\\ &= \pi E(N(0,1)^2) + (1-\pi)E(N(\mu,1)^2) && (9) \end{aligned}$$

We know the variance of a random variable X is

$$\text{var}(X) = E(X^2) - E(X)^2 \tag{10}$$

Both normal random variables in (9) have variance 1. Applying (10) to (9), we have

$$E(Y^2 \mid \mu, \pi) = \pi(1 + 0^2) + (1 - \pi)(1 + \mu^2) = 1 + (1 - \pi)\mu^2 \tag{11}$$

□

2. One type of experiment of microarray technology is summarized as follows:

   (a) A chip has a very large number of spots, say, $i = 1, \cdots, n = 10000$ spots. In each different specific spot i we place a different specific gene i.

   (a) In each spot i, approximately equal abundance of cancerous and non-cancerous tissue are then placed. We can then compare the relative abundance of gene i's DNA that binds to the cancerous versus to the non-cancerous tissue. Basically, for each gene i, this comparison leads to constructing a statistic Y_i.

   For this problem set, assume that (a) a gene i is either equally expressed in both types of tissue (in which case we say $Z_i = 1$) or it is not (in which case we say that $Z_i = 0$); (b) $Z_i, i = 1, \cdots, 10000$ are i.i.d. Bernoulli trials with probability $\text{pr}(Z = 1) = \pi$; (c) Y_i will be a draw from $N(0, 1)$ if gene i is equally expressed in both types of tissue, but Y_i will be a draw from $N(\mu, 1)$ (with $\mu \neq 0$) if gene i is not equally exressed in the two types of tissue; (d) (Z_i, Y_i) are independent vectors *across* different genes i; (e) we observe Y_i but do not know Z_i, π or μ. We are interested in estimating $(1 - \pi)$, the proportion of genes not equally expressed, and ultimately in studying further those genes that we think are not equally expressed.

   (i) You are sent by email 10000 observations Y_i of a chip. By using problem 1 above, find moment estimates of the proportion $1 - \pi$ of genes that are not equally expressed, and for μ.

   Proof. From (4) and (11) we have

$$E(Y^2 \mid \mu, \pi) = 1 + E(Y \mid \mu, \pi)\mu \tag{12}$$

$$\frac{E(Y^2 \mid \mu, \pi) - 1}{E(Y \mid \mu, \pi)} = \mu \tag{13}$$

   and then from (4) and (13),

$$E(Y \mid \mu, \pi) = (1 - \pi)\frac{E(Y^2 \mid \mu, \pi) - 1}{E(Y \mid \mu, \pi)} \tag{14}$$

$$\frac{E(Y \mid \mu, \pi)^2}{E(Y^2 \mid \mu, \pi) - 1} = 1 - \pi \tag{15}$$

□

(v) Find the maximum likelihood estimate (MLE) of $1-\pi$ and μ (e.g., use `optim` in R; in the likelihood function, parametrize π by its logit; use starting values equal to the moment estimates). If we allow π to be in $[0,1]$, are all the conditions in (A.1) to (A.4) in p. 516 of the text satisfied? Discuss whether we should worry or not worry about this issue with our data.

Proof. The likelihood function for Y is the same as (5). Thus

$$L(\theta; y) = L(\mu, 1-\pi; y) = \prod f_Y(y) \tag{16}$$

And the log-likelihood function is

$$\ell(\theta; y) = \log(L(\theta; y)) = \sum \log(\pi) + \log(\varphi_{0,1}(y)) + \log(1-\pi) + \log(\varphi_{\mu,1}(y)) \tag{17}$$

We want to maximize the log-likelihood over the parameter space. Since we maximize numerically, we want the parameters to take values in $\mathbb{R}$. Hence we map $\pi \in (0,1)$ to $\log(\pi/1-\pi) = x$. The inverse mapping is $x \mapsto \frac{e^x}{1+e^x} = \pi$ for $x \in \mathbb{R}$. With this parameterization, we maximize

$$\sum \log\left(\frac{e^x}{1+e^x}\right) + \log(\varphi_{0,1}(y)) + \log\left(\frac{1}{1+e^x}\right) + \log(\varphi_{\mu,1}(y)) \tag{18}$$

over $x, \mu \in \mathbb{R}$. This gives MLEs for x, μ. By the invariance of MLE, using the inverse mapping for x, we get the MLE for π, and hence can get the MLE for $1-\pi$. □

(vi) Find an approximate 95% confidence interval for $1-\pi$ based on the MLEs in part (v).

Proof. By the course slides, we know for a function of parameters, $g(\theta)$, the MLE $g(\hat{\theta}_n)$ has

$$\sqrt{n}[g(\hat{\theta}_n) - g(\theta)] \xrightarrow{d} N\left(0, \frac{\partial g(\theta)}{\partial \theta}' I^{-1}(\theta) \frac{\partial g(\theta)}{\partial \theta}\right) \tag{19}$$

Let $g(\theta) = g((1-\pi, \mu)) = 1-\pi$. It follows that $\nabla g((1-\pi, \mu)) = (1, 0)$.

Consider the log-likelihood in (17). Then the scores for observation i are

$$S_{i,1-\pi} = \frac{\partial \ell}{\partial(1-\pi)} = \frac{-1}{1-(1-\pi)} + \frac{1}{1-\pi} = \frac{2\pi - 1}{\pi(1-\pi)} \tag{20}$$

and

$$S_{i,\mu} = \frac{\partial \ell}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\frac{1}{\sqrt{2\pi}} e^{-(y_i-\mu)^2/2}\right) = \frac{1}{\sqrt{2\pi}} e^{-(y_i-\mu)^2/2}(y_i - \mu) \tag{21}$$

We know that the information is

$$I(\theta) = I((1-\pi, \mu)) = E\left[\begin{pmatrix} S_{i,1-\pi} \\ S_{i,\mu} \end{pmatrix} \begin{pmatrix} S_{i,1-\pi} & S_{i,\mu} \end{pmatrix}\right] \tag{22}$$

For this problem, we calculate

$$\left[\begin{pmatrix} S_{i,1-\pi} \\ S_{i,\mu} \end{pmatrix} \begin{pmatrix} S_{i,1-\pi} & S_{i,\mu} \end{pmatrix}\right] \tag{23}$$

for each observation Y_i. Then we take the mean of all those results to get $\overline{I(\theta)}$. Denote the inverse of $\overline{I(\theta)}$ as $\overline{I^{-1}(\theta)}$. Note,

$$\frac{\partial g(\theta)'}{\partial \theta}\overline{I^{-1}(\theta)}\frac{\partial g(\theta)}{\partial \theta} = \begin{pmatrix} 1 & 0 \end{pmatrix} \overline{I^{-1}(\theta)} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = [\overline{I^{-1}(\theta)}]_{1,1} \tag{24}$$

Hence, to calculate a 95% confidence interval,

$$0.95 \approx P(-1.96\sqrt{[\overline{I^{-1}(\theta)}]_{1,1}} \leq \sqrt{n}((1-\pi^{\text{MLE}}) - (1-\pi)) \leq 1.96\sqrt{[\overline{I^{-1}(\theta)}]_{1,1}}) \tag{25}$$

$$= P(1.96\sqrt{[\overline{I^{-1}(\theta)}]_{1,1}}n^{-1/2} + 1 - \pi^{\text{MLE}} \geq 1-\pi \geq -1.96\sqrt{[\overline{I^{-1}(\theta)}]_{1,1}}n^{-1/2} + 1 - \pi^{\text{MLE}}) \tag{26}$$

The calculations in R give

$$\overline{I(\theta)} = \begin{pmatrix} 200.2009 & -1.5267 \\ -1.5267 & 0.0278 \end{pmatrix} \quad \text{and} \quad \overline{I^{-1}(\theta)} = \begin{pmatrix} 0.0086 & 0.4706 \\ 0.4706 & 61.7120 \end{pmatrix} \tag{27}$$

And a confidence interval for $1-\pi$, using (26), is

$$[0.0639, 0.0675] \tag{28}$$

QUESTIONS

- How do you approximate $I(\theta)$? My thought is to use the MLE for $1-\pi$ and μ, calculate a score for each observation, average them, and that is a good estimate of $I(\theta)$. But why do you use the MLE? Does that estimate converge to $I(\theta)$ because of the central limit theorem? Are there other convergence theorems involved here that I do not see?
- Is (22) correct? Where does that come from? Is it problematic that (23) is not invertible?
- Is it correct that (20) does not depend on the observation? I believe the math, but why should that be so?

□

$$(\beta-\beta_2)((\beta-\beta_2)y + (\beta-\beta_1)p) + (\beta-\beta_1)((\beta-\beta_2)p + (\beta-\beta_1)x) \tag{29}$$