James K. Pringle
140.674 Stat Theory
Dr. Constantine Frangakis
Assignment 6
May 12, 2014

Assignment 6

The microarray experiment

1. Let Z be a Bernoulli random variable with probability $\text{pr}(Z = 1) = \pi$; 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. Let ϕ be the density of a standard normal distribution. Then

   $$\text{pr}(Y = y \mid Z = z, \mu, \pi) = \phi(y)^z \phi(y - \mu)^{1-z}$$

   Therefore,

   $$\begin{aligned} \text{pr}(Y = y \mid \mu, \pi) &= \int \text{pr}(Y = y \mid Z = z, \mu, \pi)\text{pr}(Z = z \mid \mu, \pi) dz \\ &= \int \phi(y)^z \phi(y - \mu)^{1-z} \pi^z (1 - \pi)^{1-z} dz \\ &= \phi(y)\pi + \phi(y - \mu)(1 - \pi) \end{aligned}$$

   Finally, calculating,

   $$\begin{aligned} E(Y \mid \mu, \pi) &= \int y(\phi(y)\pi + \phi(y - \mu)(1 - \pi)) dy \\ &= \pi \int y\phi(y) dy + (1 - \pi) \int y\phi(y - \mu) dy \\ &= (1 - \pi)\mu \end{aligned}$$

   □

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

Proof. Calculating,

$$\begin{aligned} E(Y^2 \mid \mu, \pi) &= \int y^2(\phi(y)\pi + \phi(y-\mu)(1-\pi))dy \\ &= \pi \int y^2\phi(y)dy + (1-\pi)\int y^2\phi(y-\mu)dy \\ &= \pi + (1-\pi)(1+\mu^2) \\ &= 1 + (1-\pi)\mu^2 \end{aligned}$$

□

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 <u>across</u> 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 (i) and (ii) above, we have

$$\begin{aligned} E(Y^2 \mid \mu, \pi) &= 1 + [(1-\pi)\mu]\mu \\ &= 1 + E(Y \mid \mu, \pi)\mu \\ \frac{E(Y^2 \mid \mu, \pi) - 1}{E(Y \mid \mu, \pi)} &= \mu \end{aligned}$$

and then from (i) and the solution for μ above,

$$\begin{aligned} E(Y \mid \mu, \pi) &= (1-\pi)\mu \\ &= (1-\pi)\frac{E(Y^2 \mid \mu, \pi) - 1}{E(Y \mid \mu, \pi)} \\ \frac{E(Y \mid \mu, \pi)^2}{E(Y^2 \mid \mu, \pi) - 1} &= 1 - \pi \end{aligned}$$

The appropriate method of moments estimates are

$$E(Y \mid \mu, \pi) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y} \quad \text{and}$$

$$E(Y^2 \mid \mu, \pi) = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 = \overline{Y^2}$$

Using the method of moments estimates in the equations above, estimates for the parameters of interest are

$$\mu^{MOM} = \frac{\overline{Y^2} - 1}{\overline{Y}} \quad \text{and}$$

$$(1-\pi)^{MOM} = \frac{\overline{Y}^2}{\overline{Y^2} - 1}$$

□

(ii) Assuming π is in $(0, 1)$ but in a set bounded away from 0 or 1, find the large sample distribution of the estimator for $\log(\pi/(1-\pi))$ resulting from (i). Hence, report an approximate 95% CI for the proportion $1 - \pi$ of genes that are not equally expressed.

Proof. Notice from question 2, part (i),

$$\begin{aligned} \frac{E(Y \mid \mu, \pi)^2}{E(Y^2 \mid \mu, \pi) - 1} &= 1 - \pi \\ \pi &= 1 - \frac{E(Y \mid \mu, \pi)^2}{E(Y^2 \mid \mu, \pi) - 1} \\ \pi &= \frac{E(Y^2 \mid \mu, \pi) - E(Y \mid \mu, \pi)^2 - 1}{E(Y^2 \mid \mu, \pi) - 1} \end{aligned}$$

Thus

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

Simplify notation by setting $\mu_1 = E(Y \mid \mu, \pi)$ and $\mu_2 = E(Y^2 \mid \mu, \pi)$. Then

$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{E(Y^2 \mid \mu, \pi) - E(Y \mid \mu, \pi)^2 - 1}{E(Y \mid \mu, \pi)^2}\right) = \log\left(\frac{\mu_2 - \mu_1^2 - 1}{\mu_1^2}\right)$$

Then the estimator for $\log \frac{\pi}{1-\pi}$ using the method of moments is

$$\log\left(\frac{\pi}{1-\pi}\right)^{MOM} = \log\left(\frac{\overline{Y^2} - \overline{Y}^2 - 1}{\overline{Y}^2}\right)$$

Define

$$g(x, y) = \log\left(\frac{x - y^2 - 1}{y^2}\right)$$

Thus

$$\frac{\partial g}{\partial x} = \frac{1}{x - y^2 - 1}$$
$$\frac{\partial g}{\partial y} = \frac{-2(x-1)}{y^3}\frac{y^2}{x - y^2 - 1} = \frac{2(x-1)}{y^3 - xy + y}$$

By the Central Limit Theorem,

$$\sqrt{n}\left(\begin{bmatrix}\overline{Y^2} \\ \overline{Y}\end{bmatrix} - \begin{bmatrix}\mu_2 \\ \mu_1\end{bmatrix}\right) \xrightarrow{D} N(0, \Sigma)$$

where

$$\Sigma = \begin{bmatrix}\text{var}(Y^2) & \text{cov}(Y^2, Y) \\ \text{cov}(Y, Y^2) & \text{var}(Y)\end{bmatrix}$$

Calculating,

$$\begin{aligned}\sqrt{n}(g(\overline{Y^2}, \overline{Y}) - g(\mu_2, \mu_1)) &= \sqrt{n}\left(\log\left(\frac{\overline{Y^2} - \overline{Y}^2 - 1}{\overline{Y}^2}\right) - \log\left(\frac{\mu_2 - \mu_1^2 - 1}{\mu_1^2}\right)\right) \\ &= \sqrt{n}\left(\log\left(\frac{\overline{Y^2} - \overline{Y}^2 - 1}{\overline{Y}^2}\right) - \log\left(\frac{\pi}{1-\pi}\right)\right)\end{aligned}$$

Other calculations give,

$$\nabla g(\mu_2, \mu_1)^T = \left[\frac{1}{\mu_2 - \mu_1^2 - 1},\ \frac{2(\mu_2 - 1)}{\mu_1^3 - \mu_2\mu_1 + \mu_1}\right]$$

By the Continuous Mapping Theorem, a consistent estimator for $\nabla g(\mu_2, \mu_1)^T$ is

$$\nabla g(\overline{Y^2}, \overline{Y})^T = \left[\frac{1}{\overline{Y^2} - \overline{Y}^2 - 1},\ \frac{2(\overline{Y^2} - 1)}{\overline{Y}^3 - (\overline{Y^2})\overline{Y} + \overline{Y})}\right]$$

Calculating, ¡¡¿¿= g1 ¡- 1/(y2$_b$ar$-y_b$ar$*y_b$ar$-1)g2 < -2*(y2_b$ar$-1)/($y_bar*y_bar*$ $y_b$ar$-y2_b$ar$*y_b$ar$+y_b$ar$)gradient_h$at$< -c(g1, g2)gradient_h$at@$Furthermore, aconsistentestim$ is

$$\hat{\Sigma} = \begin{bmatrix}\frac{1}{n-1}\sum_{i=1}^n (Y_i^2 - \overline{Y^2}) & \frac{1}{n-1}\sum_{i=1}^n (Y_i^2 - \overline{Y^2})(Y_i - \overline{Y}) \\ \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})(Y_i^2 - \overline{Y^2}) & \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})\end{bmatrix}$$

Calculating, ¡¡¿¿= s11 ¡- var(y$_s$q)s21 < $-cov(y, y_sq)s12 < -cov(y_sq, y)s22 < -var(y)sigma_hat < -matrix(c(s11, s21, s12, s22), nrow = 2, ncol = 2)@Thus, bytheDeltaMeti N(0, \nabla g(\mu_2, \mu_1)^T \Sigma \nabla g(\mu_2, \mu_1))$

By Slutsky's Theorem, a consistent estimator for $\nabla g(\mu_2, \mu_1)^T \Sigma \nabla g(\mu_2, \mu_1)$ is $\hat{\Gamma} := \nabla g(\overline{Y^2}, \overline{Y})^T \hat{\Sigma} \nabla g(\overline{Y^2}, \overline{Y})$. Calculating, ¡¡¿¿= gamma$_h$at < $-drop(gradient_hatgamma_hat@ThusbySlu N(0, 1)$Thus an approximate 95% confidence interval for $\log \frac{\pi}{1-\pi}$ is

$$\left[\log\left(\frac{\overline{Y^2} - \overline{Y}^2 - 1}{\overline{Y}^2} \right) - 1.96\sqrt{\frac{\hat{\Gamma}}{n}},\ \log\left(\frac{\overline{Y^2} - \overline{Y}^2 - 1}{\overline{Y}^2} \right) + 1.96\sqrt{\frac{\hat{\Gamma}}{n}} \right]$$

Since

$$1 - \frac{\exp\left(\log \frac{\pi}{1-\pi}\right)}{1 + \exp\left(\log \frac{\pi}{1-\pi}\right)} = 1 - \frac{\frac{\pi}{1-\pi}}{1 + \frac{\pi}{1-\pi}} = 1 - \frac{\frac{\pi}{1-\pi}}{\frac{1}{1-\pi}} = 1 - \pi$$

then applying the function

$$h(x) = 1 - \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(x)}$$

to the confidence interval above will give a 95% confidence interval for $1 - \pi$. The desired confidence interval is

$$\left[\left(1 + \frac{(\overline{Y^2} - \overline{Y}^2 - 1) \exp\left\{ 1.96\sqrt{\frac{\hat{\Gamma}}{n}} \right\}}{\overline{Y}^2} \right)^{-1},\ \left(1 + \frac{(\overline{Y^2} - \overline{Y}^2 - 1) \exp\left\{ -1.96\sqrt{\frac{\hat{\Gamma}}{n}} \right\}}{\overline{Y}^2} \right)^{-1} \right]$$

Calculating, ¡¡¿¿= t1 ¡- (y2$_b$ar $- y_bar * y_bar - 1)/(y_bar * y_bar) left_bound < -1/(1 + t1 * exp(qnorm(0.975) * sqrt(gamma_hat/n)))right_bound < -1/(1+t1*exp(qnorm(0.025) * sqrt(gamma_hat/n)))c(left_bound, right_bound)@Henceanapproximate95\%confidenceintervalfor1-$ π is

$$\begin{bmatrix} left_bound & right_bound \end{bmatrix}$$

□

Allowing π to be in [0, 1], can your CI of part (ii) be taken as evidence that there are some genes that are differentially expressed?

Proof. The above confidence interval cannot contain 0 by construction (the function $h(x)$ is positive). This confidence interval is not appropriate for testing $1 - \pi = 0$, or equivalently, $\pi = 1$, that none of the genes are differentially expressed. □

State one (alternative ?) way of testing whether or not there are some genes that are differentially expressed. Report the result from this test.

Proof. An alternative way to test that some genes are differentially expressed is to test if each gene came from the standard normal distribution. This needs multiple testing correction, since there are $n = 10000$ tests and the null hypothesis is rejected if one of those tests is significant. If the significance value is $p_0 = 0.05$, then using the Bonferroni correction, the significant p-value under multiple testing is $p_0/n = 5e-6$. ¡¡¿¿= p0 ¡- 0.05 p$_c ritical <$ $-p0/nleft2 < -qnorm(p_critical)right2 < -qnorm(1-p_critical)c(left2, right2)@Thus, anyY_i$ values outside of $[left2, right2]$ would be considered differentially expressed under this hypothesis test. Those values are ¡¡¿¿= y[which(y ¿ right2 — y ¡ left2)] @ □

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

$$L(\theta; y) = L(\mu, \pi; y) = \prod_{i=1}^{n} \text{pr}(y \mid \mu, \pi)$$

Thus, the log-likelihood function is

$$\ell(\theta; y) = \log(L(\theta; y)) = \sum_{i=1}^{n} \log(\phi(y)\pi + \phi(y-\mu)(1-\pi))$$

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

$$\sum_{i=1}^{n} \log\left(\phi(y)\frac{e^x}{1+e^x} + \phi(y-\mu)\frac{1}{1+e^x}\right) \tag{1}$$

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 $1-\pi$. ¡¡¿¿= Log likelihood function get$_l og_l ik <- function(param, y = y)t1 <- dnorm(y) * exp(param[2])/(1 + exp(param[2]))t2 <- dn$

opt$_o ut <- optim(c(mu_m om, log(pi_m om/(1-pi_m om))), get_l og_l ik, y = y, control = list(fnscale = -1), hessian = TRUE)mu_h at <- opt_o ut$par[1] pi$_h at <- exp(opt_o ut$par[2])/(1+exp(opt$_o ut$par[2])) c(mu$_h at, pi_h at)@Hence, the MLE sare$

Next, we check the conditions for the consistency of MLEs found on pg. 516 of Casella and Berger. Do the conditions hold?

(A1) **Yes**, all the Y_i are i.i.d. and the density is $\phi(y)\pi + \phi(y-\mu)(1-\pi)$.

(A2) **Yes**, the parameters are identifiable. If $(\mu, \pi) \neq (\mu', \pi')$, then $\phi(y)\pi + \phi(y-\mu)(1-\pi) \neq \phi(y)\pi' + \phi(y-\mu')(1-\pi')$ unless $\mu = \mu' = 0$. However, by definition, the genes are differentially expressed, so μ or μ' cannot be 0 in the context of this problem.

(A3) **Almost**. The densities have common support ($y \in \mathbb{R}$), however, the densities are differentiable in all of the parameter space except where $\pi = 1$ and $\pi = 0$ because they are on the boundary of the allowable values $[0, 1]$. It may be possible to define the derivative at $\pi = 1$ and $\pi = 0$ as the limit of the derivatives as $\pi \uparrow 1$ and $\pi \downarrow 0$, respectively.

(A4) **No**, the true parameter may not be an element of an open set contained in the parameter space. That is because if $\pi = 0$ or $\pi = 1$, then π is on the boundary, and there is no open set containing the boundary contained in the parameter space.

However, by hypothesis, we assume that $\pi \neq 0$ and $\pi \neq 1$. That is, we assume that there are some genes that are differentially expressed, but not all of them are differentially expressed. Therefore, if it is assumed that $\pi \in (0, 1)$, then the concerns in (A3) and (A4) are resolved. We should not worry about those issues with our data. □

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

Proof. Let $\theta_0 = (\mu_0, \text{logit}(\pi_0))$, the true values of the parameters, and let $\hat{\theta} = (\hat{\mu}, \text{logit}(\hat{\pi}))$, the MLE. The information of one data point is

$$I_1(\theta_0) = -E\left(\frac{\partial^2}{\partial\theta^2} \log \text{pr}(Y \mid \theta) \mid \theta\right)$$

By the Law of Large Numbers, and by large sample theory for MLE, a consistent estimator of the information of one data point is

$$\hat{I}_1(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log \text{pr}(y_i \mid \hat{\theta})$$

According to large sample theory for the MLE,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} N(0, I_1(\theta_0)^{-1})$$

Therefore, by Slutsky's theorem,

$$\sqrt{n}\hat{I}_1(\theta_0)^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, I)$$

where $I_1(\theta_0)^{1/2}$ is the Cholesky decomposition of $I_1(\theta)$. The output of `optim` returns the Hessian at the MLE, i.e.

$$H(\hat{\theta}) = \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log \text{pr}(y_i \mid \hat{\theta})$$

Notice

$$\sqrt{n}\hat{I}_1(\theta_0)^{1/2} = \left(-\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log \text{pr}(y_i \mid \hat{\theta})\right)^{1/2} = (-H(\hat{\theta}))^{1/2}$$

Therefore,

$$(-H(\hat{\theta}))^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, I)$$

and for the purposes of creating confidence intervals,

$$\hat{\theta} - \theta_0 \approx N(0, (-H(\hat{\theta}))^{-1})$$

Thus, an approximate 95% confidence interval for $\log \frac{\pi}{1-\pi}$ is

$$\left[\text{logit}\,(\hat{\pi}) - 1.96\sqrt{[(-H(\hat{\theta}))^{-1}]_{22}},\ \text{logit}\,(\hat{\pi}) + 1.96\sqrt{(-H(\hat{\theta}))^{-1}]_{22}}\right]$$

and applying $h(x)$, a 95% confidence interval for $1 - \pi$ is

$$LB = \left(1 + \exp\left\{\text{logit}\,(\hat{\pi}) + 1.96\sqrt{[(-H(\hat{\theta}))^{-1}]_{22}}\right\}\right)^{-1}$$
$$UB = \left(1 + \exp\left\{\text{logit}\,(\hat{\pi}) - 1.96\sqrt{[(-H(\hat{\theta}))^{-1}]_{22}}\right\}\right)^{-1}$$

where LB and UB are the lower bound and upper bound, respectively. □

We will set a cutoff c, so that for all genes with $Y_i > c$, we will proceed to do further experiments. Define the positive predictive value $PPV(c, \mu, \pi)$ to be the fraction, among the genes we get with this cutoff, that are truly differentially expressed ($Z_i = 0$), and define the negative predictive value $NPV(c, \mu, \pi)$ to be the fraction, among the genes we leave out of this cutoff, that are equally expressed ($Z_i = 1$). Use the MLEs for the parameters μ, π. Among all pairs of $(PPV(c), NPV(c))$ with a $PPV > 90\%$, what should c be to give the highest NPV? Hence state which genes from the data set you would pick based on that c.

Proof. Recall

$$\begin{aligned} PPV(c) &= P(Z_i = 0 \mid Y_i > c) \\ &= \frac{P(Y_i > c \mid Z_i = 0)P(Z_i = 0)}{P(Y_i > c \mid Z_i = 0)P(Z_i = 0) + P(Y_i > c \mid Z_i = 1)P(Z_i = 1)} \\ &= \frac{\Phi(\mu - c)(1 - \pi)}{\Phi(\mu - c)(1 - \pi) + \Phi(-c)\pi} \end{aligned}$$

and

$$\begin{aligned} NPV(c) &= P(Z_i = 1 \mid Y_i \leq c) \\ &= \frac{P(Y_i \leq c \mid Z_i = 1)P(Z_i = 1)}{P(Y_i \leq c \mid Z_i = 1)P(Z_i = 1) + P(Y_i \leq c \mid Z_i = 0)P(Z_i = 0)} \\ &= \frac{\Phi(c)\pi}{\Phi(c)\pi + \Phi(c - \mu)(1 - \pi)} \end{aligned}$$

It is clear that PPV is increasing in c and NPV is decreasing in c. Thus, the minimum c such that $PPV \geq 0.9$ gives the highest NPV among the c such that $PPV > 0.9$. That value is

$\text{PPV}(c^*) =$

$\text{NPV}(c^*) =$

□