James K. Pringle
140.673
Dr. Constantine Frangakis
Assignment 1
7 February 2013, Thursday

Problems 1, 2, 3, 4, 5

(1) Among the population, P, of women who visit phyisicians for screening for a disease, assume that the screening test has specificity and sensitivity as the one discussed in lecture one, where, here, probability statements mean fractions of women in the population P (for example, specificity of 98% means that, of all true negative women in P, 98% would test negative.). For a woman i, denote $\theta(i) = 1$ if the woman is truly positive, and 2 if truly negative; and denote $(l_i(a_1), l_i(a_2))$ to be the loss to that woman if treated, or if not treated, respectively (in the last expressions, her true status is captured already in the notation "i"). Suppose that the averages of the losses in the diseased and non-diseased women, if treated and if not treated, are:

$$\begin{aligned} l(1, a_1) &:= E\{l_i(a_1)|\theta(i) = 1\} = 2 \\ l(1, a_2) &:= E\{l_i(a_2)|\theta(i) = 1\} = 5 \\ l(2, a_1) &:= E\{l_i(a_1)|\theta(i) = 2\} = 1 \\ l(2, a_2) &:= E\{l_i(a_2)|\theta(i) = 2\} = 0 \end{aligned}$$

but that l_i may generally vary from woman to woman, and that among women of a particular status (diseased, or not diseased), the losses l_i may be correlated with the value X_i that the diagnostic test would show for woman i. For the strategy s defined as $s(X_i) = a_1$ if X_i is positive, and $s(X_i) = a_2$ if X_i is negative, which of the following conditions 1.-3. would make the losses

$$E\{l_i(s(X_i))|\theta(i)\} \text{ and } E\{l(\theta(i), s(X_i))|\theta(i)\} \tag{1}$$

equal and why?

1. For a fixed action a, $l_i(a)$ is constant within women of common disease status $\theta(i)$.
2. For a fixed action a, X_i is independent of $l_i(a)$.
3. For a fixed action a, X_i is independent of $l_i(a)$ given $\theta(i)$.

Proof. Let's examine 1. For an action a, the loss $l_i(a)$ is constant within women of common disease status $\theta(i)$. Since, for example, $E\{l_i(a_1)|\theta(i) = 1\} = 2$, it must be that $l_i(a_1)$ given $\theta(i) = 1$ is 2 for all i. Otherwise, if $l_i(a_1)$ were some other value, $c \neq 2$, then the expected value would be c. In other words, we have $l_i(s(X_i)) = l(\theta(i), s(X_i))$, since $s(X_i)$ is the same on both sides of the equal sign and $\theta(i)$ is captured in l_i. Since we have that equality, taking expectations over subsets of people with true status $\theta(i)$ will yield equal values.

Now 2. And 3. □

(2) Now assume that the way the test X_i is determined is by measuring a continuous variable X_i^*, and calling X_i positive if $X^* > 0$, otherwise calling X_i negative. Assuming that $\text{pr}(X_i^*|\theta(i))$ is normal with variance 1, find $E(X_i^*|\theta(i))$ for the two disease conditions. Also, assume that, for a fixed action a, $\text{pr}(l_i(a)|\theta(i))$ is normal with the means given above, variance 10, and that $cor(l_i(a), X_i^*|\theta(i)) = 0.7$. Using simulation of 1000 diseased and 1000 non-diseased women, or otherwise, estimate the two average losses in (1).

Proof. From the lecture notes, the specificiy of the test is 0.98 and the sensitivity is 0.94. Thus $P(X_i^* > 0|\theta(i) = 1) = 0.94$. Since we assume a normal distribution, with variance $\sigma^2 = 1$, we have

$$\begin{aligned} 0.94 &= P(X^* > 0|\theta(i) = 1) \\ 0.94 &= P((X^* - \mu_1)/\sigma > -\mu_1/\sigma|\theta(i) = 1) \\ 0.94 &= P(Z > -\mu_1|\theta(i) = 1) \end{aligned}$$

Thus we solve for $\mu_1 = E(X_i^*|\theta(i) = 1)$ in $1 - F_Z(-\mu_1) = 0.94$ where F_Z is the distribution function for the standard normal distribution. It follows that $\mu_1 = -F_Z^{-1}(0.06) = -\texttt{qnorm}(0.06) = 1.554774$.

Similarly solving for $\mu_2 = E(X_i^*|\theta(i) = 2)$, one finds that

$$\begin{aligned} 0.98 &= P(X^* < 0|\theta(i) = 2) \\ 0.98 &= P((X^* - \mu_2)\ \sigma < -\mu_2/\sigma|\theta(i) = 2) \\ 0.98 &= P(Z < -\mu_2|\theta(i) = 2) \end{aligned}$$

So $\mu = -F_Z^{-1}(0.98) = -\texttt{qnorm}(0.98) = -2.053749$.

The simulations yielded 2.184648 for the first expectation and 2.156 for the second expectation. □

(3) Assume that the random variable X has finite $E|X|$ and is continuous (has a density). Show that $E|X - a|$ is minimun at $a = \text{median}(X)$.

Proof. Let $f(x)$ be the density function of X. Calculating,

$$\begin{aligned} \frac{d}{da}E|X| &= \frac{d}{da}\int_{-\infty}^{\infty} |x - a|f(x)dx \\ &= \int_{-\infty}^{\infty} \frac{d}{da}|x - a|f(x)dx \\ &= \int_{-\infty}^{a} f(x)dx + \int_{a}^{\infty} -f(x)dx \\ &= \int_{-\infty}^{a} f(x)dx - \int_{a}^{\infty} f(x)dx \end{aligned}$$

Setting this derivative to 0, we have equality for the value of a for which

$$F(a) = \int_{-\infty}^{a} f(x)dx = \int_{a}^{\infty} f(x)dx = 1 - F(a).$$

That is where $1 - F(a) = F(a) = 0.5$, or when $a = \text{median}(X)$. The second derivate with respect to a is $d/da(F(a) - (1 - F(a))) = 2f(a) > 0$. Thus the a that we found is a minimum for the original function. □

(4) We want to estimate the true value of the scalar θ, and we have a loss function $l(\theta, a) = |\theta - a|$. Based on previous similar studies, we believe that, a priori, $\mathcal{P}(\theta) = N(\mu_0, \tau_0^2)$, where μ_0 and τ_0 are known values. To help us estimate θ, we design a study that gives us data X where $\mathcal{P}(X|\theta) = N(\theta, \sigma_0^2)$, and where σ_0^2 is assumed known.

1. Find the posterior distribution $\mathcal{P}(\theta|X)$.

   Proof. According to `http://zoe.bme.gatech.edu/~bv20/bmestat/Bank/bayes.pdf`, the posterior distribution is

$$\theta|X \sim N\left(\frac{\tau_0^2}{\sigma_0^2 + \tau_0^2}X + \frac{\sigma_0^2}{\sigma_0^2 + \tau_0^2}\mu_0, \frac{\sigma_0^2\tau_0^2}{\sigma_0^2 + \tau_0^2}\right)$$

   □

2. Using Exercise 3, find the Bayes estimator for this problem, i.e. the estimate $s(X)$ that minimizes $E\{E(l(\theta, s(X))|\theta)\}$, where the outer expectation is with respect to the prior distribution for θ.

(5) Refer to Problem 4, and suppose we have iid observations from the likelihood. By considering a sequence of priors, each as in problem 4, but with mean 0 and τ_0 increasing with the sequence, show that the sample average is a minimax estimator. (Hint: note that the sample average is equalizer).