Problem Set 1.

1. Among the population, P, of women who visit physicians for screening for a disease, assume that the screening test has specificity and sensitivity as the one discussed in lecture 1, 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:

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

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

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

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 $\theta(i)$.

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 pr$(X_i^*|\theta(i))$ is normal with variance 1, find $E(X_i^*|\theta(i))$ for the two disease conditions. Also, assume that, conditionally on $\theta(i)$, the variables $l_i(a_1), l_i(a_2), X_i^*$ are jointly normal, and that, for a fixed action a, pr$(l_i(a) \mid \theta(i))$ has the means given above, variance 10, and that $cor(l_i(a), X_i^*|\theta(i))$=.7. Using simulation of 1000 diseased and 1000 non diseased women, or otherwise, estimate the two average losses in (1).

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=median(X).

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, pr$(\theta) = N(\mu_0, \tau_0^2)$, where μ_0 and τ_0^2 are known values. To help us estimate θ, we design a study that gives us data X where pr$(X \mid \theta) = N(\theta, \sigma_0^2)$, and where σ_0^2 is assumed known.

1. Find the posterior distribution $\text{pr}(\theta \mid X)$.

2. Using Exercise 3, find the Bayes estimator for this problem, i.e., the estimator $s(X)$ that minimizes $E\{E(l(\theta, s(X)) \mid \theta)\}$, 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).