Probability and Statistics
Supervision 2

Samuel Lee

Question 1

Two parameter points θ and θ' in a parameter set Θ are said to be observationally equivalent if $L(\theta|\mathcal{X}^n) = L(\theta'|\mathcal{X}^n)$ for all samples $\mathcal{X}^n$ consisting of $X_i, i \in \{1, \ldots, n\}$. Conversely, a parameter point θ in Θ is said to be identifiable if there is no other $\theta' \in \Theta$ with $\theta' \neq \theta$ that is observationally equivalent to it.

For a uniform distribution $U[\theta_0, \theta_1]$, we have

$$\begin{aligned} L(\theta_0, \theta_1|\mathcal{X}^n) &= \prod_{X_i \in [\theta_0, \theta_1]} \left(\frac{1}{\theta_1 - \theta_0}\right) \prod_{X_i \notin [\theta_0, \theta_1]} 0 \\ &= \begin{cases} \frac{1}{(\theta_1 - \theta_0)^n} & \text{if } X_i \in [\theta_0, \theta_1] \forall i \in \{1, \ldots, n\} \\ 0 & \text{if } \exists i : X_i \notin [\theta_0, \theta_1] \end{cases} \end{aligned}$$

It is apparent that for any pair of θ_0, θ_1 and sample $\mathcal{X}^n$, there exists some sample $\mathcal{X'}^n$ which can deliver a different value of the likelihood function. For example, if all the observations in $\mathcal{X}^n$ are in the interval $[\theta_0, \theta_1]$, the likelihood of θ_0, θ_1 is $\frac{1}{(\theta_1 - \theta_0)^n}$. Any sample $\mathcal{X'}^n$ where a single observation is not in $[\theta_0, \theta_1]$ obviously yields a likelihood of 0 for θ_0, θ_1, since it would be literally impossible to have any observations outside of that interval if θ_0 and θ_1 were correctly identified. Thus, θ_0 and θ_1 are identifiable.

This also applies to a case where we only have one observation. It only takes one aberrant observation to invalidate any given pair of θ_0, θ_1, so these parameters are still identifiable.

Question 2

(a)

The marginal distribution of Y is

$$\Pr(Y = y) = \begin{cases} (1 - p_X)(1 - p_{Y0}) + p_X(1 - p_{Y1}), & y = 0 \\ (1 - p_X)p_{Y0} + p_X p_{Y1}, & y = 1 \\ 0, & \text{otherwise} \end{cases}$$

(b)

The joint distribution of X, Y is

$$\Pr(X = x, Y = x) = \begin{cases} (1-p_X)(1-p_{Y0}), & x = 0, y = 0 \\ p_X(1-p_{Y1}), & x = 1, y = 0 \\ (1-p_X)p_{Y0}, & x = 0, y = 1 \\ p_X p_{Y1}, & x = 1, y = 1 \\ 0, & \text{otherwise} \end{cases}$$

(c)

Given a sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, the likelihood function for the data is

$$\prod_{i=1}^{n} \Pr(X = X_i, Y = Y_i)$$

(d)

The estimator $\hat{p}_X$ is $\frac{1}{n}\sum_{i=1}^{n} 1(X_i = 1)$ which is equivalent to $\frac{1}{n}\sum X_i$. It is an unbiased estimator of p_X, since $\mathrm{E}[\hat{p}_X] = \frac{1}{n}\sum \mathrm{E}[X_i] = p_X$. It can be shown that the variance of the estimator is $\frac{p_X(1-p_X)}{n}$. It is also the maximum likelihood estimator of p_X. To show this, we note that the log-likelihood for the sample obtained is

$$\ell(p_X, p_Y | \mathcal{X}^n) = k \ln p_X + (n-k)\ln(1-p_X) + R(p_{Y0}, p_{Y1})$$

where k is the number of times $X = 1$ and R is some remainder containing only p_{Y0} and p_{Y1}. Maximizing this yields

$$\frac{k}{\hat{p}_X} - \frac{n-k}{1-\hat{p}_X} = 0$$
$$\hat{p}_X = \frac{k}{n} = \frac{1}{n}\sum_{i=1}^{n} 1(X_i = 1)$$

Likewise, the expected value of $\hat{p}_Y$ is $\Pr(Y = 1) = (1-p_X)p_{Y0} + p_X p_{Y1}$, and its variance is $\frac{1}{n}[(1-p_X)p_{Y0} + p_X p_{Y1}][1 - (1-p_X)p_{Y0} - p_X p_{Y1}]$.

(e)

We could use the method of moments to estimate p_{Y0} and p_{Y1}. In the population,

$$\mathrm{E}[Y|X = 0] = p_{Y0}$$
$$\mathrm{E}[Y|X = 1] = p_{Y1}$$

So we could just estimate the parameters by

$$\hat{p}_{Y0} = \frac{1}{n(1-\hat{p}_X)}\sum_{i=1}^{n} 1(X_i = 0, Y_i = 1)$$
$$\hat{p}_{Y1} = \frac{1}{n\hat{p}_X}\sum_{i=1}^{n} 1(X_i = 1, Y_i = 1)$$

In effect, we are splitting the sample into two groups, one where $X = 0$ and the other where $X = 1$, and then calculating $\hat{p}_Y$ for those two subsamples separately. $n(1 - \hat{p}_X)$ is the number of observations where $X = 0$ and $n\hat{p}_X$ is the number of observations where $X = 1$.

Question 3

With $X_i \sim N(\theta, \theta)$, $\hat{\theta}_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, and $\hat{\theta}_2 = s^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, we have

$$\hat{\theta}(\alpha) = \alpha\hat{\theta}_1 + (1-\alpha)\hat{\theta}_2 = \frac{\alpha}{n}\sum_{i=1}^n X_i + \frac{1-\alpha}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

which is also equal to

$$\begin{aligned}
\hat{\theta}(\alpha) &= \frac{\alpha}{n}\sum_{i=1}^n X_i + \frac{1-\alpha}{n}\sum_{i=1}^n (X_i - \theta - (\bar{X} - \theta))^2 \\
&= \frac{a}{n}\sum_{i=1}^n X_i + \frac{1-\alpha}{n}\sum_{i=1}^n (X_i - \theta)^2 + \frac{1-\alpha}{n}\sum_{i=1}^n (\bar{X} - \theta)^2 - 2\frac{1-\alpha}{n}\sum_{i=1}^n (X_i - \theta)(\bar{X} - \theta) \\
&= \frac{a}{n}\sum_{i=1}^n X_i + \frac{1-\alpha}{n}\sum_{i=1}^n (X_i - \theta)^2 - \frac{1-\alpha}{n}\sum_{i=1}^n (\bar{X} - \theta)^2 \\
&= \frac{a}{n}\sum_{i=1}^n X_i + \frac{1-\alpha}{n}\sum_{i=1}^n (X_i - \theta)^2 - (1-\alpha)(\bar{X} - \theta)^2 \\
&= \frac{\alpha}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i + \frac{1-\alpha}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \theta)^2 - (1-\alpha)(\bar{X} - \theta)^2 \\
&= \frac{\alpha}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \theta) + \frac{1-\alpha}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^n [(X_i - \theta)^2 - \theta] + \theta - (1-\alpha)(\bar{X} - \theta)^2
\end{aligned}$$

and therefore

$$\sqrt{n}[\hat{\theta}(\alpha) - \theta] = \alpha \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \theta)}_{\xrightarrow{D} N(0,\theta)} + (1-\alpha) \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n [(X_i - \theta)^2 - \theta]}_{\xrightarrow{D} N(0, \mathrm{Var}[(X_i - \theta)^2])} - \underbrace{\sqrt{n}(1-\alpha)(\bar{X} - \theta)^2}_{\xrightarrow{P} 0}$$

which converges in distribution to $N(0, V)$ where

$$V = \alpha^2\theta + (1-\alpha)^2 \mathrm{Var}[(X_i - \theta)^2] + 2\alpha(1-\alpha)\mathrm{Cov}[X_i, (X_i - \theta)^2]$$

The last expression is equal to zero since

$$\begin{aligned}
\mathrm{Cov}[X_i, (X_i - \theta)^2] &= \mathrm{Cov}[X_i, X_i^2 - 2\theta X_i + \theta^2] \\
&= \mathrm{Cov}[X_i, X_i^2] - 2\theta\mathrm{Cov}[X_i, X_i] \\
&= 0
\end{aligned}$$

where it can be shown that $\mathrm{Cov}[X_i, X_i^2] = 2\mu\sigma^2 = 2\theta^2$ and $\mathrm{Cov}[X_i, X_i] = \sigma^2 = \theta$. Therefore,

$$\sqrt{n}[\hat{\theta}(\alpha) - \theta] \xrightarrow{D} N(0, V), \ V = \alpha^2\theta + (1-\alpha)^2\mathrm{Var}[(X_i - \theta)^2]$$

To minimize the variance of the asymptotic distribution, we note that

$$\frac{\partial V}{\partial \alpha} = 2\alpha\theta - 2(1-\alpha)\mathrm{Var}[(X_i-\theta)^2] \implies \frac{\partial^2 V}{\partial \alpha^2} = 2\theta + 2\mathrm{Var}[(X_i-\theta)^2] > 0$$

since $\theta > 0$ must be true for θ to be a valid quantity for the variance of X_i. This means that the optimal α can be found by setting $\dfrac{\partial V}{\partial \alpha} = 0$:

$$2\alpha^*\theta - 2\mathrm{Var}[(X_i-\theta)^2] + 2\alpha^*\mathrm{Var}[(X_i-\theta)^2] = 0 \implies \alpha^* = \frac{\mathrm{Var}[(X_i-\theta)^2]}{\mathrm{Var}[(X_i-\theta)^2]+\theta}$$

Actually, with some messy work that won't be shown here, and using $\mathrm{E}[X^2] = \mu^2+\sigma^2, \mathrm{E}[X^3] = \mu^3+3\mu\sigma^2, \mathrm{E}[X^4] = \mu^4+6\mu^2\sigma^2+3\sigma^4$ for $X \sim N(\mu,\sigma^2)$, it can be shown that $\mathrm{Var}[(X_i-\theta)^2] = 2\theta^2$. Therefore,

$$\sqrt{n}[\hat{\theta}(\alpha)-\theta] \xrightarrow{d} N\left(0, \alpha^2\theta + 2(1-\alpha)^2\theta^2\right)$$

and $\alpha^* = \frac{2\theta}{2\theta+1}$.

Question 4

If it were true that $\mu_X = 0$, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ would have a $N\left(0,\frac{1}{n}\right)$ distribution. If this were the case, $\Pr(\sqrt{n}\bar{X} \leq x) = \Phi(x)$ and $\Pr(|\sqrt{n}\bar{X}| \geq 1.96) \approx 0.05$. So we can calculate $\sqrt{n}\bar{X}$, and if its absolute value exceeds 1.96, we conclude that this would only happen with a 5% probability if μ_X were truly 0 and we reject the hypothesis that $\mu_X = 0$ at a 5% significance level.

Assuming we also did this with Y, we would have the t-statistic $\sqrt{n}\bar{Y}$ as well. However, rejecting the hypotheses that $\mu_X = 0$ and $\mu_Y = 0$ separately is not sufficient for us to reject $\mu_X = \mu_Y = 0$ at the same significance level, because

$$\begin{aligned}
&\Pr\left(|\sqrt{n}\bar{X}| \geq 1.96 \cup |\sqrt{n}\bar{Y}| \geq 1.96 \middle| \mu_X = \mu_Y = 0\right) \\
=&\Pr\left(|\sqrt{n}\bar{X}| \geq 1.96 \middle| \mu_X = 0\right) + \Pr\left(|\sqrt{n}\bar{Y}| \geq 1.96 \middle| \mu_Y = 0\right) \\
&- \Pr\left(|\sqrt{n}\bar{X}| \geq 1.96 \cap |\sqrt{n}\bar{Y}| \geq 1.96 \middle| \mu_X = \mu_Y = 0\right) \\
=&0.05 + 0.05 - (0.05)^2 = 0.0975 > 0.05
\end{aligned}$$

However, $n\bar{X}^2$ and $n\bar{Y}^2$ are $\chi^2(1)$ distributed, and therefore their sum $n(\bar{X}^2+\bar{Y}^2)$ is $\chi^2(2)$ distributed. We can do a test at the 5% level using that test statistic instead with the relevant rejection criterion.

Question 5

We assume that the tanks are randomly captured from the population. When $k = 1$, this is analogous to taking a sample from a discrete uniform distribution $U\{1, N\}$. The expected value of a single sample X is

$$\begin{aligned}\mathrm{E}[X] &= \sum_{i=1}^{N} \Pr(X = i) \times i \\ &= \frac{1}{N}\sum_{i=1}^{N} i \\ &= \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2}\end{aligned}$$

When $k = 1$, a possible estimator could be $\hat{N} = 2X_1 - 1$. $\hat{N}$ is unbiased since

$$\mathrm{E}[\hat{N}] = \mathrm{E}[2X_i - 1] = 2\mathrm{E}[X_i] - 1 = N + 1 - 1 = N$$

When $k = 2$, we have two serial numbers, one of which is larger than the other. Hypothetically, we could just commit ourselves to discard the first or second observation without looking at it, and use whatever's left to calculate the unbiased estimate using the estimator derived before. However, if we happen to throw away the larger number, this could lead to an estimate that is actually lower than the number we threw away. In this case, information has been wasted by potentially throwing away the larger number.

Before any tanks have been captured, it is equally likely for the larger number to be on the first or second tank. Therefore, we can commit ourselves to focus on maximum number, M, and there is no relevant information wasted in doing so. In other words, M is a sufficient statistic. (We know everything about the population distribution except the maximum value; in particular, we know that the serial numbers are spaced apart in steps of 1, so the exact value of the smaller serial number doesn't tell us anything we don't already know.) The expected value of M is

$$\begin{aligned}\mathrm{E}[M] &= \sum_{i=1}^{N} \Pr(M = i) \times i \\ \mathrm{E}[M] &= \sum_{i=1}^{N} [\Pr(X_1 = i, X_2 < i) + \Pr(X_1 < i, X_2 = i)] \times i \\ &= \sum_{i=1}^{N} \left[\frac{1}{N} \times \frac{i-1}{N-1} + \frac{i-1}{N} \times \frac{1}{N-1}\right] \times i \\ &= \frac{2}{N(N-1)} \sum_{i=1}^{N} (i^2 - i) \\ &= \frac{2}{N(N-1)} \left(\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)}{2}\right) \\ &= \frac{2}{N(N-1)} \left(\frac{N(N+1)(N-1)}{3}\right) \\ &= \frac{2}{3}(N+1)\end{aligned}$$

and thus, our unbiased estimator can be $\hat{N} = \frac{3}{2}M - 1$.

Question 6

With $P(d) = \log_{10}(d+1) - \log_{10}(d)$, $d \in \{1, \ldots, 9\}$,

$$\Pr(D \leq d) = \sum_{i=1}^{d} \log_{10}(i+1) - \sum_{i=1}^{d} \log_{10}(i) = \sum_{i=1}^{d} \log_{10}(i+1) - \sum_{i=0}^{d-1} \log_{10}(i+1) = \log_{10}(d+1)$$

Given some dataset, we'd extract the leading significant digits $\mathcal{D}^n = \{D_1, \ldots, D_n\} \in \{1, \ldots, 9\}^n$, and ideally find a test statistic $T(\mathcal{X}^n)$ and rejection region $\mathcal{R}$ such that (at the 5% significance level)

$$\Pr\left(T(\mathcal{D}^n) \in \mathcal{R} \middle| \Pr(D_i \leq d) = \log_{10}(d+1)\right) = 0.05$$

If we take it that the hypothesis is true, and D_i is independently and identically distributed with $\Pr(D_i = d) = \log_{10}(d+1) - \log_{10}(d)$, then $D_i \sim M(n, P(1), \ldots, P(9))$ where M denotes the multinomial distribution.