STAT3005 Assignment 7
Lam Hoi Chun (1155192755)

Question 1

$$
\begin{aligned}
\hat{\mu} &= \int_{-\infty}^{\infty} x\hat{f}(x)dx \\
&= \int_{-\infty}^{\infty} x \cdot \frac{1}{nl}\sum_{i=1}^{n} K\left(\frac{X_i - x}{l}\right) dx \\
&= \frac{1}{nl}\sum_{i=1}^{n} \int_{-\infty}^{\infty} x \cdot K\left(\frac{X_i - x}{l}\right) dx
\end{aligned}
$$

Consider the integral, by letting $u = (X_i - x)/l$

$$
\begin{aligned}
\int_{-\infty}^{\infty} x \cdot K\left(\frac{X_i - x}{l}\right) dx &= l\int_{-\infty}^{\infty} (X_i - lu)K(u)du \\
&= l\left[X_i \int_{-\infty}^{\infty} K(u)du - l\int_{-\infty}^{\infty} u \cdot K(u)du\right] \\
&= l(X_i \cdot 1 - l \cdot 0) \\
&= lX_i
\end{aligned}
$$

So we have

$$
\begin{aligned}
\hat{\mu} &= \frac{1}{nl}\sum_{i=1}^{n}(lX_i) \\
&= \frac{nl\bar{X}}{nl} \\
&= \bar{X}
\end{aligned}
$$

Question 2

$$\begin{aligned}\hat{\sigma}^2 &= \int_{-\infty}^{\infty} x^2 \cdot \hat{f}(x) dx - \hat{\mu}^2 \\ &= \frac{1}{nl} \sum_{i=1}^{n} \int_{-\infty}^{\infty} x^2 \cdot K\left(\frac{X_i - x}{l}\right) dx\end{aligned}$$

Similar to Q1, by letting $u = (X_i - x)/l$

$$\begin{aligned}\int_{-\infty}^{\infty} x^2 \cdot K\left(\frac{X_i - x}{l}\right) dx &= l \int_{-\infty}^{\infty} (X_i - lu)^2 K(u) du \\ &= l \int_{-\infty}^{\infty} (X_i^2 - 2X_i lu - l^2 u^2) K(u) du \\ &= l \left[X_i^2 \int_{-\infty}^{\infty} K(u) du - 2X_i l \int_{-\infty}^{\infty} u \cdot K(u) du - l^2 \int_{-\infty}^{\infty} u^2 \cdot K(u) du \right] \\ &= l \left[X_i^2 \cdot 1 - 2X_i l \cdot 0 - l^2 \cdot \sigma_K^2 \right] \\ &= lX_i^2 + l^3 \sigma_K^2\end{aligned}$$

So we have

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{nl} \sum_{i=1}^{n} (lX_i^2 + l^3 \sigma_K^2) - \hat{\mu}^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} X_i^2 + \frac{1}{nl}(nl^3 \sigma_K^2) - \hat{\mu}^2 \\ &= \overline{X^2} + l^2 \sigma_K^2 - \overline{X}^2 \\ &= \overline{X^2} - \overline{X}^2 + l^2 \sigma_K^2\end{aligned}$$

Question 3

When $l \to 0, \hat{\sigma}^2 \to \overline{X^2} - \overline{X}^2$. The estimator approach to the sample variance, which is an unbiased estimator for σ^2 for small l, so it is a good estimator.

Question 4(a)

The plot of the KDE $\tilde{f}_1$ of f_1 and the KDE $\tilde{f}_0$ of f_0 can be generated using the following R code:

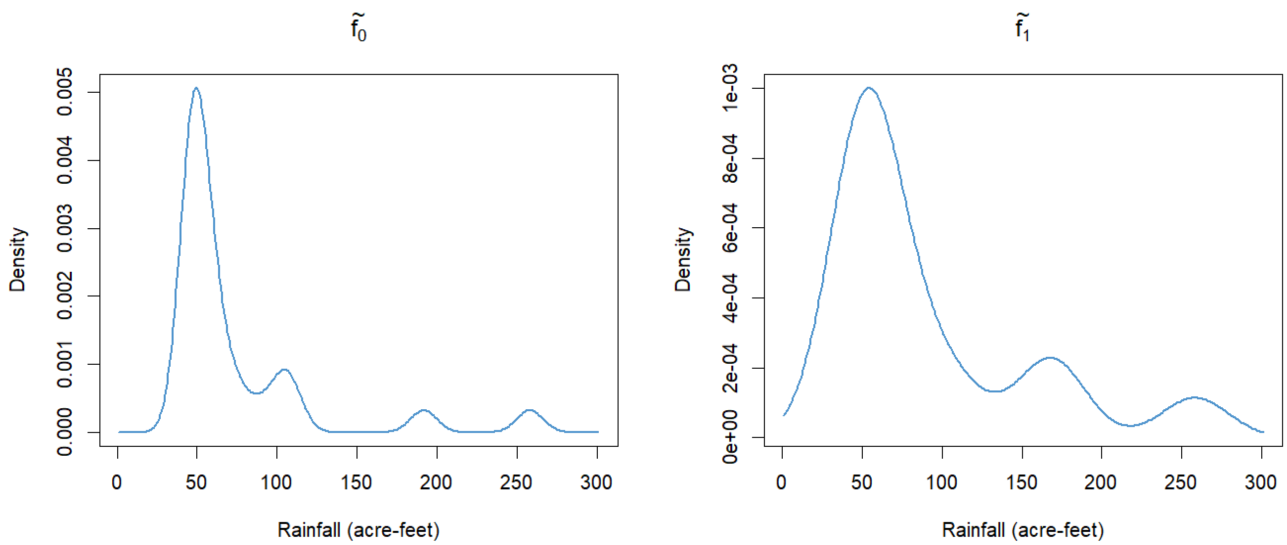
```r
data = read.csv("clouds.csv")
A = data$A
X = data$X

not_seeded <- data$X[data$A == 0]
seeded <- data$X[data$A == 1]

tilde_f0 = density0(not_seeded, C=0.9, plot=TRUE)$density
tilde_f1 = density0(seeded, C=0.9, plot=TRUE)$density

par(mfrow=c(1,2))
plot(tilde_f0, xlab = "Rainfall (acre-feet)", ylab = "Density",
     col = "steelblue3", type="l", lwd="1.75", main=expression(tilde(f[0])))
plot(tilde_f1, xlab = "Rainfall (acre-feet)", ylab = "Density",
     col = "steelblue3", type="l", lwd="1.75",main=expression(tilde(f[1])))
```

And the output is



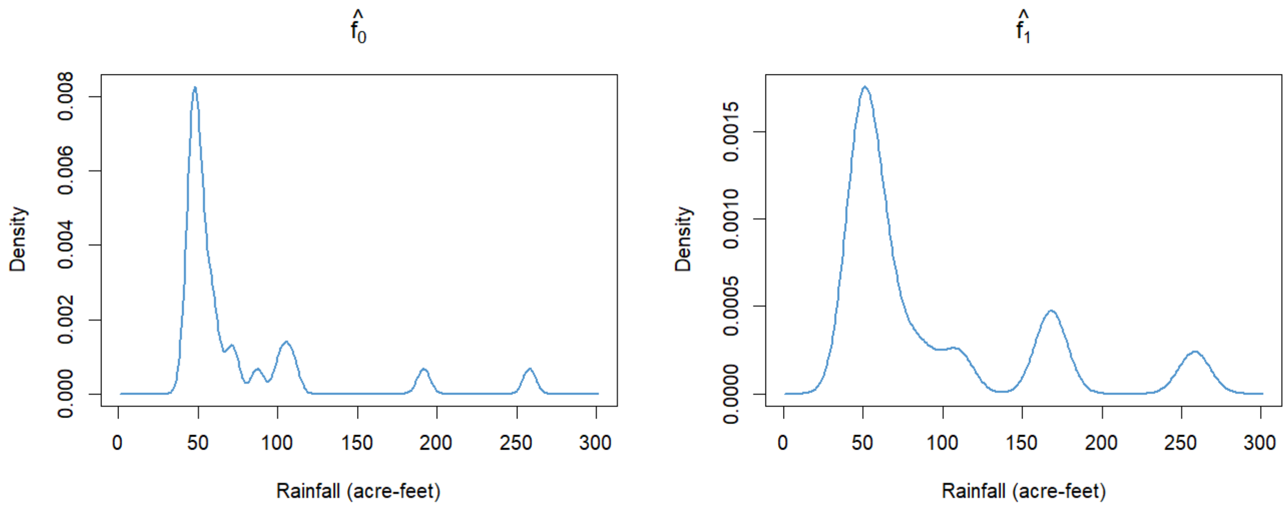
Question 4(b)

Note that the plug-in bandwidth estimator is preferred, then we can redo (a) using the following R code:

```r
hat_f0 = density0(not_seeded, bw=bw.SJ(not_seeded), plot=TRUE)$density
hat_f1 = density0(seeded, bw=bw.SJ(seeded), plot=TRUE)$density

plot(hat_f0, xlab = "Rainfall (acre-feet)", ylab = "Density",
     col = "steelblue3", type="l", lwd="1.75", main=expression(hat(f[0])))
plot(hat_f1, xlab = "Rainfall (acre-feet)", ylab = "Density",
     col = "steelblue3", type="l", lwd="1.75",main=expression(hat(f[1])))
```

And the output is



Question 4(c)

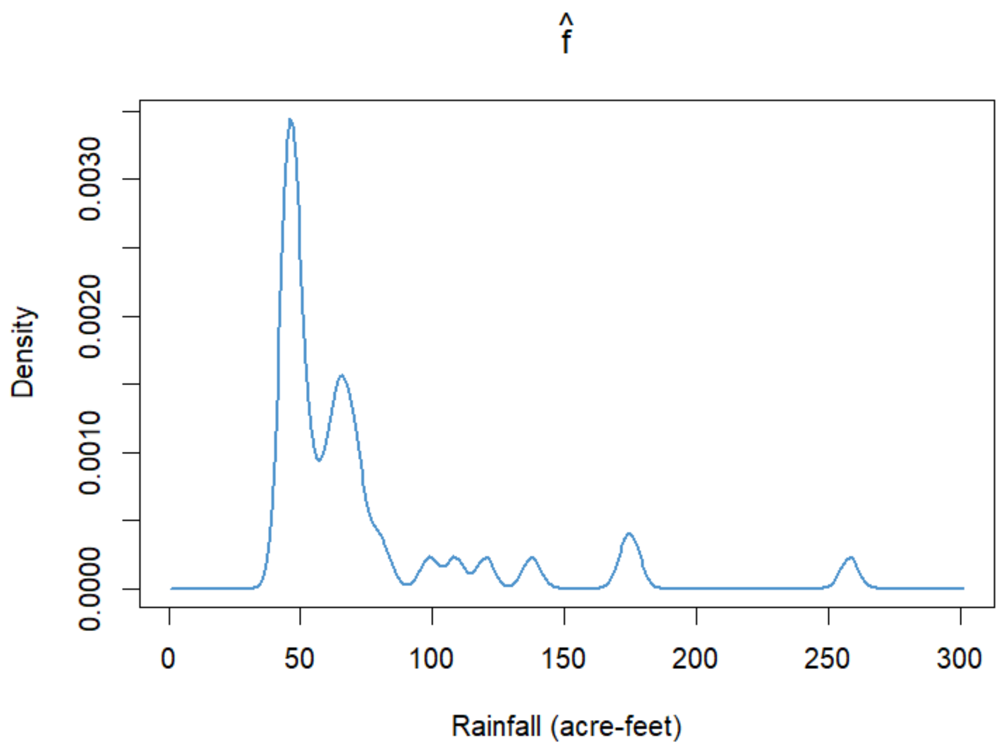
$\hat{f}_0$ and $\hat{f}_1$ have similar shapes, which both median are around 50. Meanwhile, $\hat{f}_1$ has a heavier right tail, which its density for "150+ Rainfall (acre-feet)" is much larger than that of $\hat{f}_0$.

Question 4(d)

The plot of the plug-in estimate $\hat{f}$ of f can be generated using the following R code:

```r
par(mfrow=c(1,1))
hat_f = density0(X, bw=bw.SJ(X), plot=TRUE)$density
plot(hat_f, xlab = "Rainfall (acre-feet)", ylab = "Density",
      col = "steelblue3", type="l", lwd="1.75", main=expression(hat(f)))
```

And the output is



$\hat{f}$ mix the features of $\hat{f}_0$ and $\hat{f}_1$, which has high density around the median ≈ 50. Besides, it has heavy right tail for "150+ Rainfall (acre-feet)"

Question 4(e)

$\overline{f}(x)$ is sensible as it constructs the population density from the mixture of seeded and unseeded distributions weighted by their probability.

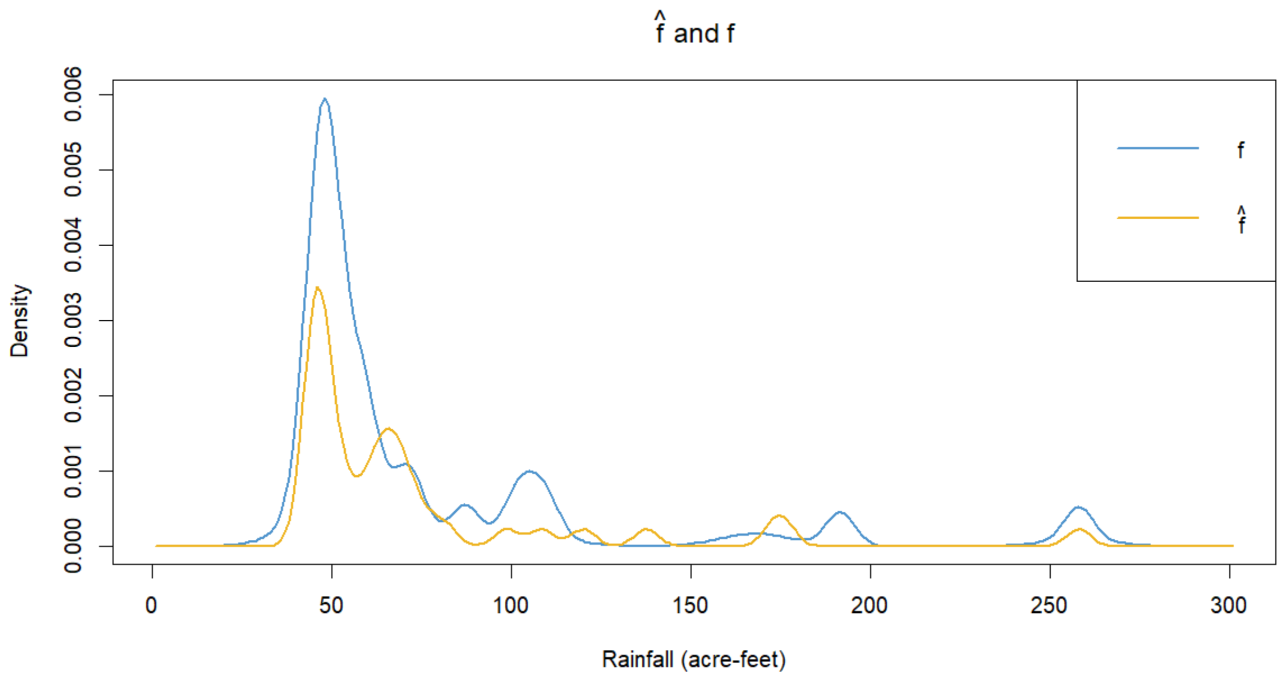
Question 4(f)

The plot can be generated by the following R code:

```r
hat_pi <- mean(data$A)
bar_f <- hat_pi * hat_f1 + (1 - hat_pi) * hat_f0

plot(bar_f, xlab = "Rainfall (acre-feet)", ylab = "Density",
      col = "steelblue3", type="l", lwd=1.75, main = expression(hat(f)~and~f))
lines(hat_f, col = "goldenrod2", type="l", lwd=2)
legend("topright", legend = c("f", expression(hat(f))),
        col = c("steelblue3","goldenrod2"), lwd=2)
```

And the output is



Question 4(g)

$\overline{f}$ combines the estimated distributions for seeded and unseeded clouds weighted by their empirical probability. Meanwhile f is only a direct KDE from all data, treating all data as one homogeneous population. Therefore, I prefer $\overline{f}$ as it can reflect the underlying experimental conditions.