Assignment # 1

Saša Milić | 997 410 626

1. (a)

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\mu)) \qquad \text{by the Central Limit Theorem}$$

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} g'(\theta)\mathcal{N}(0, \sigma^2(\mu)) \qquad \text{by the delta method}$$

$$\implies \sigma^2(\mu)g'(\theta)^2 = 1 \qquad g'(\theta)Y \sim \mathcal{N}(0, g'(\theta)^2\sigma^2(\mu)), \text{ where } Y \sim \mathcal{N}(0, \sigma^2(\mu))$$

$$\implies g'(\mu) = \pm\frac{1}{\sigma(\mu)}$$

(b) i.

$$g'(\mu) = \pm\left(\frac{1}{\sigma^2(\mu)}\right)^{1/2} = \pm\left(\frac{1}{\mu}\right)^{1/2}$$

$$\implies g(\mu) = \pm\int \mu^{-1/2}\,\mathrm{d}\mu = \boxed{\pm 2\mu^{1/2} + c} \qquad \text{where } c \text{ is some constant.}$$

ii.

$$g'(\mu) = \pm\left(\frac{1}{\sigma^2(\mu)}\right)^{1/2} = \pm\left(\frac{1}{\mu^2}\right)^{1/2} = \pm\frac{1}{\mu}$$

$$\implies g(\mu) = \pm\int \frac{1}{\mu}\,\mathrm{d}\mu = \boxed{\pm\ln\mu + c} \qquad \text{where } c \text{ is some constant.}$$

2. (a)

$$\theta(F) = E_F\left[\frac{X}{\mu(F)}\ln\left(\frac{X}{\mu(F)}\right)\right]$$

$$\geq E_F\left[\frac{X}{\mu(F)}\right]\ln\left(E_F\left[\frac{X}{\mu(F)}\right]\right) \qquad \text{by Jensen's inequality, } g(x) = x\ln(x) \text{ is convex.}$$

$$= \frac{E_F(X)}{\mu(F)}\ln\left(\frac{E_F(X)}{\mu(F)}\right)$$

$$= \ln(1) \qquad \text{Since } \mu(F) = E_F(X).$$

$$= 0$$

(b) Let c be the income of everyone in the population. Then let X be a constant random variable with the following probability distribution:

$$f(x) = \begin{cases} 1 & x = c \\ 0 & x \neq c \end{cases}$$

Now we calculate $\theta(F)$.

$$\begin{aligned} \theta(F) &= E_F\left[\frac{X}{\mu(F)} \ln\left(\frac{X}{\mu(F)}\right)\right] \\ &= \left[\frac{c}{\mu(F)} \ln\left(\frac{c}{\mu(F)}\right)\right] f\left(X = c\right) && \text{since we have } f\left(X \neq c\right) = 0 \\ &= \left[\frac{c}{c} \ln\left(\frac{c}{c}\right)\right] && \mu(F) = E_F(X) = c \\ &= \ln\left(1\right) \\ &= 0 \end{aligned}$$

(c) We have the following probability distribution for X (the income of a member of the population):

$$f(x) = \begin{cases} \epsilon & x > a \\ 1 - \epsilon & x = 0 \end{cases}$$

We first calculate $\mu(F)$.

$$\begin{aligned} \mu(F) &= E_F(X) \\ &= \int_0^\infty x f(x)\, \mathrm{d}x \\ &= \int_0^a x f(x)\, \mathrm{d}x + \int_a^\infty x f(x)\, \mathrm{d}x \\ &= \int_a^\infty x f(x)\, \mathrm{d}x && \text{Since } x \in [0, a] \Longrightarrow x = 0. \\ &= \int_a^\infty x\epsilon\ \, \mathrm{d}x \end{aligned}$$

So, $\mu(F) \to 0$ as $\epsilon \downarrow 0$.

We now look at how $\theta(F)$ behaves as $\epsilon \downarrow 0$.

$$\begin{aligned}
\theta(F) &= E_F\left[\frac{X}{\mu(F)}\ln\left(\frac{X}{\mu(F)}\right)\right] \\
&= \frac{1}{\mu(F)}E_F\left[X\ln\left(\frac{X}{\mu(F)}\right)\right] \\
&= \frac{1}{\mu(F)}E_F\left[X\ln\left(X\right) - \ln\left(\mu(F)\right)\right] \\
&\to \infty \text{ as } \epsilon \downarrow 0 \qquad\qquad \text{As } \epsilon \downarrow 0,\ \frac{1}{\mu(F)} \to \infty, \text{ and } -\ln(\mu(F)) \to \infty.
\end{aligned}$$

(d) Let the new distribution (once everyone in the population has had their income multiplied by k) be $G = kF$.

$$\begin{aligned}
\theta(G) &= E_G\left[\frac{Y}{E_F(Y)}\ln\left(\frac{Y}{E_F(Y)}\right)\right] \\
&= E_F\left[\frac{kX}{E_F(kX)}\ln\left(\frac{kX}{E_F(kX)}\right)\right] \\
&= E_F\left[\frac{kX}{kE_F(X)}\ln\left(\frac{kX}{kE_F(X)}\right)\right] \\
&= E_F\left[\frac{X}{E_F(X)}\ln\left(\frac{X}{E_F(X)}\right)\right] \\
&= \theta(F)
\end{aligned}$$

(e) Let $\mu = \mu(F) = \mu(G)$.

$$\begin{aligned}
(1-\epsilon)\theta(F) + \epsilon\theta(G) &= (1-\epsilon)E_F\left[\frac{X}{\mu}\ln\left(\frac{X}{\mu}\right)\right] + \epsilon E_G\left[\frac{X}{\mu}\ln\left(\frac{X}{\mu}\right)\right] \\
&= (1-\epsilon)\int_{-\infty}^{\infty}\left[\frac{X}{\mu}\ln\left(\frac{X}{\mu}\right)\right]f(x)\,\mathrm{d}x + \epsilon\int_{-\infty}^{\infty}\left[\frac{X}{\mu}\ln\left(\frac{X}{\mu}\right)\right]g(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\left((1-\epsilon)\left[\frac{X}{\mu}\ln\left(\frac{X}{\mu}\right)\right]f(x) + \epsilon\left[\frac{X}{\mu}\ln\left(\frac{X}{\mu}\right)\right]g(x)\right)\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\left[\frac{X}{\mu}\ln\left(\frac{X}{\mu}\right)\right]\left((1-\epsilon)f(x) + \epsilon g(x)\right)\mathrm{d}x \\
&= E_{(1-\epsilon)F+\epsilon G}\left[\frac{X}{\mu}\ln\left(\frac{X}{\mu}\right)\right] \\
&= \theta\left((1-\epsilon)F + \epsilon G\right)
\end{aligned}$$

3. (a) Approximately $0.00698n$ data points are expected to be outliers when n data points are drawn from a standard normal distribution.

```
> iqr = qnorm(.75) - qnorm(.25)
> pnorm(qnorm(.25) - 1.5 * iqr) + 1 - pnorm(qnorm(.75) + 1.5 * iqr)
[1] 0.006976603
```

(b) We solve analytically.

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{(1+x)^2}{2} & -1 \leq x \leq 0 \\ \frac{1-(1+x)^2}{2} & 0 < x \leq 1 \\ 1 & 1 > x \end{cases}$$

We have $F^{-1}(1/4) = \sqrt{1/2} - 1 \approx -0.2928$, $F^{-1}(3/4) = 1 - \sqrt{1/2} \approx 0.2928$, and $IQR = 2(\sqrt{1/2} - 1) \approx 0.5858$. And so the probability of an outlier is

$$\begin{aligned} & F\left(-0.2928 - 1.5 \times 0.5858\right) + 1 - F\left(0.2928 + 1.5 \times 0.5858\right) \\ &= F\left(-1.1715\right) + 1 - F\left(1.1715\right) \\ &= 0 + 1 - 1 \\ &= 0 \end{aligned}$$

We expect no outliers for a triangular distribution.

(c) Approximately $0.0823n$, $0.0335n$, and $0.0188n$ data points are expected to be outliers when n data points are drawn from t distributions with 2, 5 and 10 degrees of freedom, respectively.

```
> v = c(2,5,10)
> iqr = qt(.75,v) - qt(.25,v)
> pt(qt(.25,v) - 1.5*iqr, v) + 1 - pt(qt(.75,v) + 1.5*iqr, v)
[1] 0.08233706 0.03352633 0.01881880
```

(d) As the tails of the distributions in (a) – (c) become "heavier", the proportion of outliers increases, which should match our intuition.

4. (a) Figure 1 shows a histogram of our data. The data appears to be bimodal. The data does not appear to come from a single normal distribution, although it seems it could potentially come from a mixture of two normal distributions.

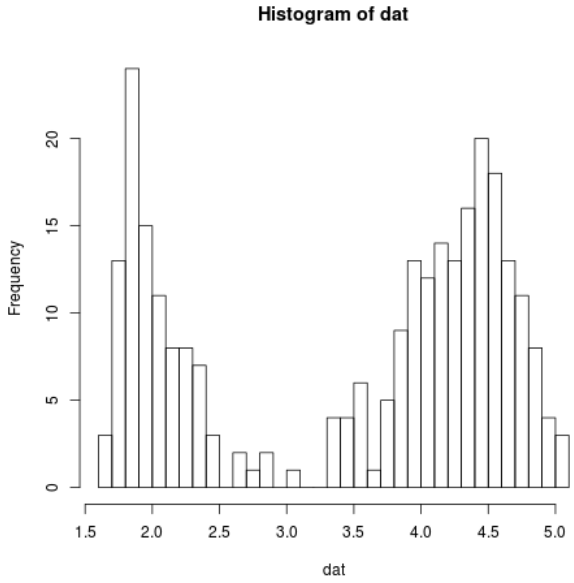


Figure 1: A histogram of 272 eruptions of the Old Faithful geyser in Yellowstone National Park.

(b) As bandwidth increases, the density plot becomes increasingly "smoother" (see Figure 2). In particular, the curves with bandwidths 0.1 and 1.7 are undersmoothed and oversmoothed, respectively.

(c) We now assume the data comes from a mixture of two normal distributions. Although not a very sophisticated method, we can look at the density plots from before (in part (b)), and use them to make "good" guesses as to the parameters of the mixture distribution. In particular, we observe that the two mounds of the distribution are centred at approximately 2 and 4.5, and both have a width of about 2, corresponding to a standard deviation of 0.5. It also appears the data is about twice as likely to lie somewhere in the rightmost mound, corresponding to a θ value of 0.67. This initial guess corresponds to values $\mu_1 = 2, \mu_2 = 4.5, \sigma_1 = 0.5, \sigma_2 = 0.5, \theta = 0.67$; the resulting density for this set of parameter estimates is shown in Figure 3a. After adjusting these parameters further, we have the density in Figure 3b, which more correctly matches the kernel density estimates obtained before.

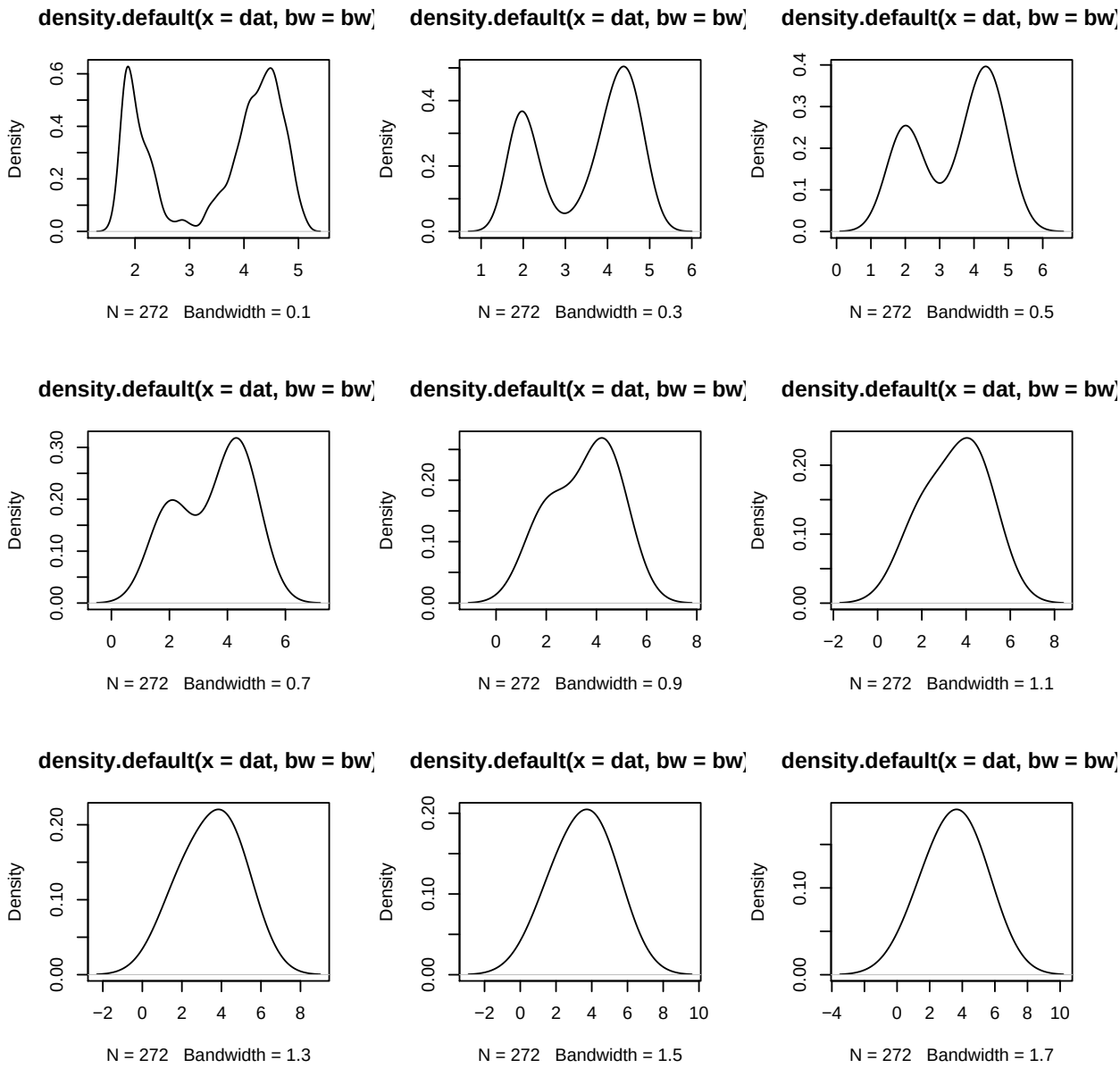
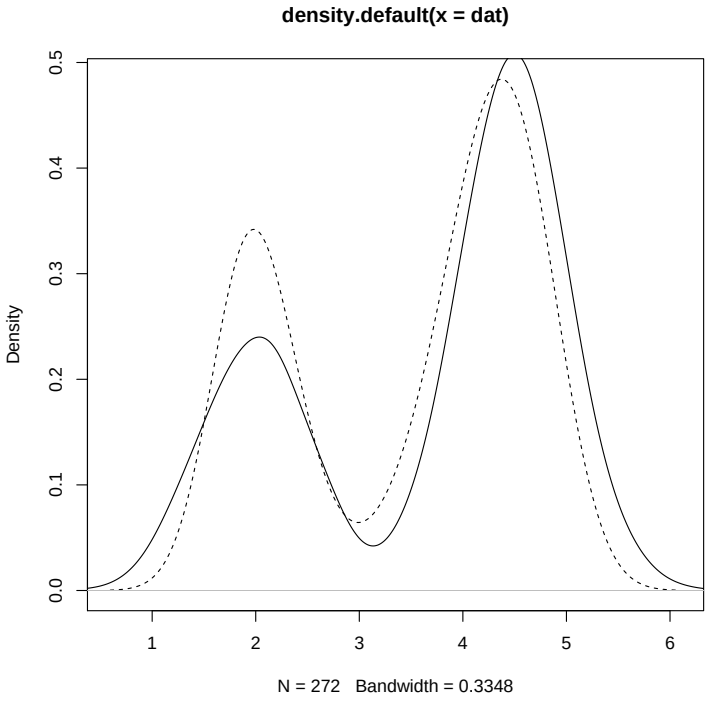
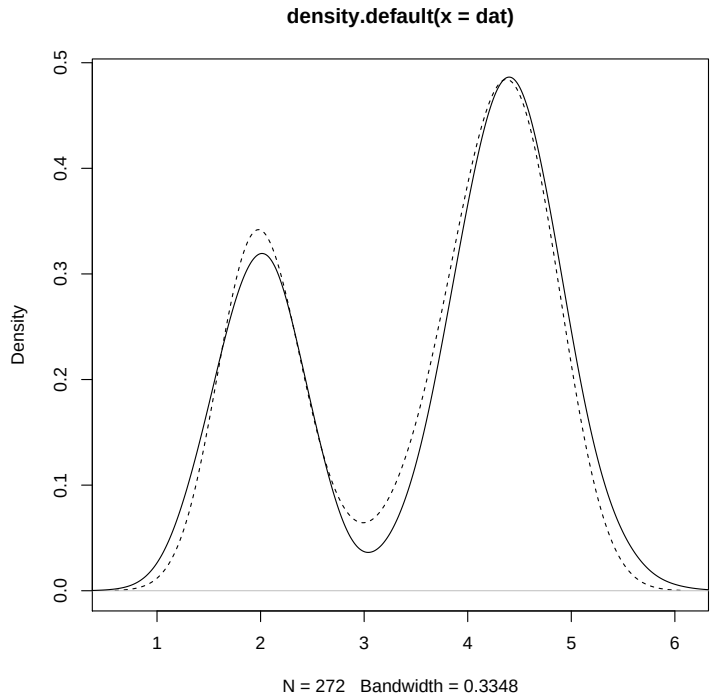


Figure 2: Estimates of the distribution of the geyser data using the R `density` function with varying bandwidths.



(a) $\mu_1 = 2, \mu_2 = 4.5, \sigma_1 = 0.5, \sigma_2 = 0.5, \theta = 0.67$



(b) $\mu_1 = 1.9, \mu_2 = 4.4, \sigma_1 = 0.4, \sigma_2 = 0.5, \theta = 0.64$

Figure 3: Estimating parameters for a mixture of two normal distributions. The kernel density estimation (with the default bandwidth in R) is the dotted line.