3602 Group Project

Mai Junshen 3035974444

Fong Chi Hong 3035923330

Chen Weiqiao 3035973581

November 2024

Contents

1 Introduction **2**

2 Project 1: True Love of Peony **2**

2.1 Background story and settings 2

2.2 Estimate 5

2.2.1 Posterior distribution 5

2.2.2 Estimation of target 8

2.3 Result and discussion 9

3 Project 2: Rather die than compromise **11**

3.1 Background story and settings 11

3.2 Maximum likelihood estimate of probability distribution 12

3.2.1 Scenario 1 12

3.2.2 Scenario 2 14

3.3 Estimation of target 15

3.3.1 Estimation of θ 15

3.3.2 Bootstrap estimation 17

3.4 Result and discussion 18

3.4.1 Scenario 1 Result 18

3.4.2 Scenario 2 Result 20

1 Introduction

This report is about our group project that extend the examples presented in the lecture by adapting our understandings and ideas on different statistical inference knowledge. Project 1 aims to propose an improved version of the Bayesian procedure to help Congwen evaluate the chance of love between him and Peony. We will use Bayesian inference and computational techniques to derive posterior distributions, enhancing our understanding of Congwen's case. Project 2 aims to use a nonparametric approach to estimate some probabilities of Yan Huizhu's experiments during her career life. We will construct the bootstrap cumulative distribution function to assess the performance of the nonparametric estimate.

2 Project 1: True Love of Peony

2.1 Background story and settings

Project 1 will evaluate how likely that the writer Shen Congwen's beloved one, Peony, loves him under the following settings.

Congwen and $\theta_1 \in \mathbb{N}^*$ other people (so-called frogs) love Peony. Peony may or may not have true love, represented by $\theta_2 \in \{-1, 0, 1, \cdots, \theta_1\}$. When $\theta_2 = -1$, Peony concentrates on studying and does not fall in love with anyone; When $\theta_2 = 0$, Peony's true love is Congwen; When $\theta_2 = j$, Peony's true love is the θ_1-th frog ($j = 1, 2, \cdots, \theta_1$).

From Day 1, every frog except Congwen, who has not sent a love letter to Peony nor received it from Peony, has a probability of $p_t := 0.2 \times 0.95^t$ (Figure 1) to send her a love letter, a decreasing function by t different from the fixed value in the original version. The frog who has sent her letter will never send another one if he does not receive any response from Peony. As for Congwen, he sends letters on Day $6, 7, 13, 14, \cdots$. If on some day, Peony receives a letter from her true love, she will reply to it with a probability θ_3, and thus with a probability of $1 - \theta_3$, she will not reply to her true love.

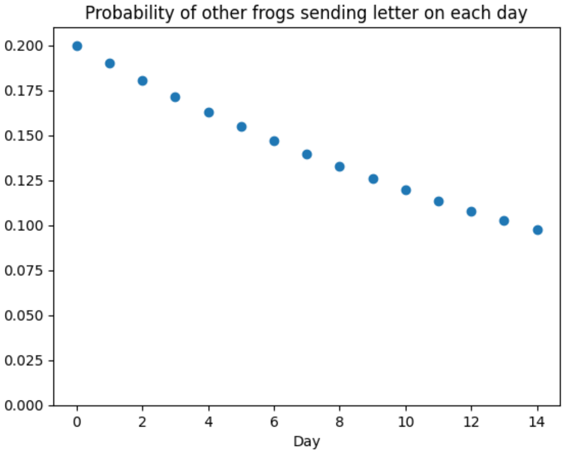


Figure 1: Plot of Probability for Other Frogs Sending Letter on Each Day

Peony can at most love one person from Congwen and other θ_1 frogs. If Peony does not love anyone, i.e., $\theta_2 = -1$, every time she receives a love letter, she may switch from loving studying to loving the sender. The probability of this caused by the letter from frog $\{1, 2, \cdots, \theta_1\}$ is a parameter θ_4, while as for Congwen's letter, it is $\gamma = 0.3$, a known value because he is confident about his letter. Assume that once Peony has fallen in love with someone, she will no longer change her mind or switch to love other. We also suppose that on days $6, 7, 13, 14, \cdots$, Peony always reads Congwen's letter first before those from the other frogs.

Data obtained:

What is known to Congwen on day t is that, from day 0 onward, there have been n_t letters sent to Peony by other frogs (shown in the table). He himself also sent m_t letters to Peony. However, Peony replies to none of the letters. We denote the indicator that Peony responds to Congwen's i-th letter as Z_i $(i \in \{1, \cdots, m_t\})$, and the indicator that Peony responds to the j-th letter from other frogs as X_j, $(j \in \{1, \cdots, n_t\})$. All the observed entries in $\mathbf{X} = [X_1, \cdots, X_{m_t}]^T$ and $\mathbf{Z} = [Z_1, \cdots, Z_{n_t}]^T$ are 0.

Table 1: Table of m_t and n_t Observed by Congwen on Each Day

Day	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
m_t	0	0	0	0	0	0	1	2	2	2	2	2	2	3	4	4
n_t	0	1	1	1	2	4	4	5	7	8	10	12	14	15	15	15

Prior distribution

We propose the prior distributions for our parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \theta_4]^T$.

1. There are $N = 100$ men acquainted with Peony and each has probability $\alpha = 0.1$ to fail in love with her, thus $\theta_1 \sim \text{Binomial}(N = 100, \alpha = 0.1)$, and $\pi(\theta_1) = \binom{N}{\theta_1}\alpha^{\theta_1}(1-\alpha)^{N-\theta_1}\mathbf{1}\{0 \le \theta_1 \le 100\}$.

2. Peony has a probability of 0.5 to love study only and has an equal chance to pick Congwen and one of the θ_1 frogs. Thus, $\text{P}(\theta_2 = -1) = 0.5$, and given θ_1, $\text{P}(\theta_2 = i) = \frac{0.5}{\theta_1 + 1}$ for $i \in \{0, 1, \cdots, \theta_1\}$.

3. $\theta_3 \sim \text{U}(0, 1)$, independent of θ_1 and θ_2. $\pi(\theta_3) = \mathbf{1}\{0 \le \theta_3 \le 1\}$

4. $\theta_4 \sim \text{Beta}(1, \beta = 8)$, independent of other parameters. $\pi(\theta_4) = \frac{\Gamma(1+\beta)}{\Gamma(\beta)}(1-\theta_4)^{\beta-1}\mathbf{1}\{0 \le \theta_4 \le 1\}$

Objective

It will be helpful to Congwen to evaluate the probability that Peony loves him on day $T \in \{0, 1, \cdots 15\}$, based on the prior distribution and the observed data $\mathbf{n} := [n_1, \cdots, n_T]^T$, his action $\mathbf{m} := [m_1, \cdots, m_T]^T$ and $\mathbf{X}, \mathbf{Z}$ being $\mathbf{0}$. Note that this includes the case that $\theta_2 = 0$ (Peony loves him at the beginning) and Peony switches to love him after reading his letter (in which case θ_2 must be -1).

It is also interesting to get the probability that Peony loves someone else or still loves studying on day T.

2.2 Estimate

This section will derive the prior and posterior distribution under our settings, and then obtain the formula of our targets in summation form.

2.2.1 Posterior distribution

We first calculate the prior distribution,

$$\begin{aligned}\pi(\boldsymbol{\theta}) =& \binom{N}{\theta_1}\alpha^{\theta_1}(1-\alpha)^{N-\theta_1}\mathbf{1}\{0 \le \theta_1 \le 100\} \times 0.5^{\mathbf{1}\{\theta_2=-1\}} \times (\frac{1-0.5}{\theta_1+1})^{\mathbf{1}\{0\le\theta_2\le\theta_1\}}\mathbf{1}\{-1 \le \theta_2 \le \theta_1\} \\ & \times \mathbf{1}\{0 \le \theta_3 \le 1\} \times \frac{\Gamma(1+\beta)}{\Gamma(\beta)}(1-\theta_4)^{\beta-1}\mathbf{1}\{0 \le \theta_4 \le 1\} \\ \propto & \frac{\alpha^{\theta_1}(1-\alpha)^{-\theta_1}}{\theta_1!(N-\theta_1)!} \times \frac{(1-\theta_4)^{\beta-1}}{(1+\theta_1)^{\mathbf{1}\{0\le\theta_2\}}}\mathbf{1}\{0 \le \theta_1 \le N, -1 \le \theta_2 \le \theta_1, 0 \le \theta_3, \theta_4 \le 1\}\end{aligned}$$

Given the parameter $\boldsymbol{\theta}$, we proceed to obtain the conditional data up to day T, namely $\mathbf{n}$ and $\mathbf{X}$, $\mathbf{Z}$ being zeros.

$$\begin{aligned}f(\mathbf{n}|\boldsymbol{\theta}) =& \mathbf{1}\{n_T \le \theta_1\}\prod_{t=1}^{T}\binom{\theta_1-n_{t-1}}{n_t-n_{t-1}} \times p_t^{n_t-n_{t-1}}(1-p_t)^{\theta_1-n_t} \\ P_{XZ} :=& P[\mathbf{X}=\mathbf{0}, \mathbf{Z}=\mathbf{0}|\mathbf{n},\boldsymbol{\theta}] = \mathbf{1}_{\theta_2>n_T} + \mathbf{1}_{\theta_2=0}(1-\theta_3)^{m_T} + \sum_{i=1}^{n_T}\mathbf{1}_{\theta_2=i}(1-\theta_3) + \mathbf{1}_{\theta_2=-1}\times \\ & \left[(1-\theta_4)^{n_T}(1-\gamma)^{m_T} + \sum_{j=1}^{m_T}(1-\theta_3)^{m_T-j+1}\gamma(1-\theta_4)^{n(j)}(1-\gamma)^{j-1}\right. \\ & \left. +(1-\theta_3)\sum_{i=1}^{n_T}\theta_4(1-\theta_4)^{i-1}(1-\gamma)^{m(i)}\right] \\ =& \begin{cases} 1 & , \quad \theta_2 > n_T \\ (1-\theta_3)^{m_T} & , \quad \theta_2 = 0 \\ 1-\theta_3 & , \quad 1 \le \theta_2 \le n_T \\ h(\theta_3,\theta_4) & , \quad \theta_2 = -1 \end{cases}\end{aligned}$$

where

$$h(\theta_3, \theta_4) := (1-\theta_4)^{n_T}(1-\gamma)^{m_T} + \sum_{j=1}^{m_T}(1-\theta_3)^{m_T-j+1}\gamma(1-\theta_4)^{n(j)}(1-\gamma)^{j-1} + (1-\theta_3)\sum_{i=1}^{n_T}\theta_4(1-\theta_4)^{i-1}(1-\gamma)^{m(i)}$$

It will be useful to calculate

$$\begin{aligned}
H(a,b) :=& \int_0^1\int_0^1 \theta_4^a(1-\theta_4)^b h(\theta_3,\theta_4)\mathrm{d}\theta_3\mathrm{d}\theta_4 \\
=& \int_0^1\int_0^1 \theta_4^a(1-\theta_4)^{n_T+b}(1-\gamma)^{m_T} + \sum_{j=1}^{m_T}(1-\theta_3)^{m_T-j+1}\theta_4^a(1-\theta_4)^{n(j)+b}\gamma(1-\gamma)^{j-1} \\
&+ (1-\theta_3)\sum_{i=1}^{n_T}\theta_4^{1+a}(1-\theta_4)^{i-1+b}(1-\gamma)^{m(i)}\mathrm{d}\theta_3\mathrm{d}\theta_4 \\
=& \int_0^1 \theta_4^a(1-\theta_4)^{n_T+b}(1-\gamma)^{m_T}\mathrm{d}\theta_4 + \sum_{j=1}^{m_T}\frac{\gamma(1-\gamma)^{j-1}}{m_T-j+1}\int_0^1\theta_4^a(1-\theta_4)^{n(j)+b}\mathrm{d}\theta_4 \\
&+ \frac{1}{2}\sum_{i=1}^{n_T}(1-\gamma)^{m(i)}\int_0^1\theta_4^{1+a}(1-\theta_4)^{i-1+b}\mathrm{d}\theta_4 \\
=& \frac{a!(n_T+b)!}{(a+n_T+b+1)!}(1-\gamma)^{m_T} + \sum_{j=1}^{m_T}\frac{\gamma(1-\gamma)^{j-1}}{m_T-j+1}\times\frac{a![n(j)+b]!}{[a+n(j)+b+1]!} \\
&+ \frac{1}{2}\sum_{i=1}^{n_T}(1-\gamma)^{m(i)}\frac{(1+a)!(i-1+b)!}{(a+b+i+1)!}
\end{aligned}$$

Then, the posterior distribution of $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \theta_4]^T$ can be calculated,
(Denote $\Theta := \{n_T \le \theta_1 \le N, -1 \le \theta_2 \le \theta_1, 0 \le \theta_3, \theta_4 \le 1\}$)

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0}) \propto&\ \pi(\boldsymbol{\theta})\times f(\mathbf{n})\times P[\mathbf{X}=\mathbf{0},\mathbf{Z}=\mathbf{0}|\mathbf{n},\boldsymbol{\theta}] \\
&\propto \mathbf{1}\{\Theta\}\frac{\alpha^{\theta_1}(1-\alpha)^{-\theta_1}}{\theta_1!(N-\theta_1)!(1+\theta_1)^{\mathbf{1}\{0\le\theta_2\}}}\prod_{t=1}^{T}\frac{(\theta_1-n_{t-1})!}{(\theta_1-n_t)!}(1-p_t)^{\theta_1}\times(1-\theta_4)^{\beta-1}P_{XZ} \\
&\propto \mathbf{1}\{\Theta\}\frac{\alpha^{\theta_1}(1-\alpha)^{-\theta_1}}{(N-\theta_1)!(\theta_1-n_T)!(1+\theta_1)^{\mathbf{1}\{0\le\theta_2\}}}\times\left[\prod_{t=1}^{T}(1-p_t)\right]^{\theta_1}\times(1-\theta_4)^{\beta-1}P_{XZ}
\end{aligned}$$

We obtain the normalizing coefficient,

$$c_T^{-1} := \sum_{\theta_2=-1}^{N}\sum_{\theta_1=\max\{\theta_2,n_T\}}^{N}\int_0^1\int_0^1\frac{\alpha^{\theta_1}(1-\alpha)^{-\theta_1}}{(N-\theta_1)!(\theta_1-n_T)!(1+\theta_1)^{\mathbf{1}\{0\le\theta_2\}}}\times\left[\prod_{t=1}^{T}(1-p_t)\right]^{\theta_1}\times(1-\theta_4)^{\beta-1}P_{XZ}\mathrm{d}\theta_3\mathrm{d}\theta_4$$

Let

$$\begin{aligned}
S(\theta_2) &= \sum_{\theta_1=\max\{\theta_2,n_T\}}^{N} \int_0^1\int_0^1 \frac{\alpha^{\theta_1}(1-\alpha)^{-\theta_1}}{(N-\theta_1)!(\theta_1-n_T)!(1+\theta_1)^{\mathbf{1}\{0\le\theta_2\}}} \times \left[\prod_{t=1}^{T}(1-p_t)\right]^{\theta_1} \times (1-\theta_4)^{\beta-1} P_{XZ} \mathrm{d}\theta_3 \mathrm{d}\theta_4 \\
&= \sum_{\theta_1=\max\{\theta_2,n_T\}}^{N} \frac{\alpha^{\theta_1}(1-\alpha)^{-\theta_1}}{(N-\theta_1)!(\theta_1-n_T)!(1+\theta_1)^{\mathbf{1}\{0\le\theta_2\}}} \times \left[\prod_{t=1}^{T}(1-p_t)\right]^{\theta_1} \times \int_0^1\int_0^1 (1-\theta_4)^{\beta-1} P_{XZ} \mathrm{d}\theta_3 \mathrm{d}\theta_4
\end{aligned}$$

Then

$$c_T = \left[\sum_{\theta_2=-1}^{N} S(\theta_2)\right]^{-1}$$

Denote $q_T := \prod_{t=1}^{T}(1-p_t)$ and $D(\theta_1) = \frac{\alpha^{\theta_1}(1-\alpha)^{-\theta_1}}{(N-\theta_1)!(\theta_1-n_T)!(1+\theta_1)} \times q_T^{\theta_1}$.

Discuss the following case

1. $\theta_2 > n_T$,

$$S(\theta_2) = \sum_{\theta_1=\theta_2}^{N} D(\theta_1) \int_0^1\int_0^1 (1-\theta_4)^{\beta-1} \mathrm{d}\theta_3 \mathrm{d}\theta_4 = \sum_{\theta_1=\theta_2}^{N} \frac{D(\theta_1)}{\beta}$$

2. $1 \le \theta_2 \le n_T$,

$$S(\theta_2) = \sum_{\theta_1=n_T}^{N} D(\theta_1) \int_0^1\int_0^1 (1-\theta_3)(1-\theta_4)^{\beta-1} \mathrm{d}\theta_3 \mathrm{d}\theta_4 = \sum_{\theta_1=\theta_2}^{N} \frac{D(\theta_1)}{2\beta}$$

3. $\theta_2 = 0$,

$$S(\theta_2) = \sum_{\theta_1=n_T}^{N} D(\theta_1) \int_0^1\int_0^1 (1-\theta_3)^{m_T}(1-\theta_4)^{\beta-1} \mathrm{d}\theta_3 \mathrm{d}\theta_4 = \sum_{\theta_1=\theta_2}^{N} \frac{D(\theta_1)}{\beta(m_T+1)}$$

4. $\theta_2 = -1$,

$$S(\theta_2) = \sum_{\theta_1=n_T}^{N} (1+\theta_1) D(\theta_1) \times \int_0^1\int_0^1 (1-\theta_4)^{\beta-1} h(\theta_3,\theta_4) \mathrm{d}\theta_3 \mathrm{d}\theta_4 = \sum_{\theta_1=n_T}^{N} (1+\theta_1) D(\theta_1) H(0,\beta-1)$$

Then,

$$c_T = \left[\sum_{\theta_2=-1}^{N} S(\theta_2)\right]^{-1} \quad \text{and } \pi(\boldsymbol{\theta}|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0}) = \mathbf{1}\{\Theta\} c_T D(\theta_1)(1+\theta_1)^{\mathbf{1}\{\theta_2=-1\}}(1-\theta_4)^{\beta-1} P_{XZ}$$

All of the above are merely finite summations and can be computed by programming.

2.2.2 Estimation of target

Our target is the events $A_T := \{\text{Peony had no true love on day } T\}$ and $B_T := \{\text{Peony loved Congwen on day } T\}$. It is also interesting to calculate $C_T := (A_T \cup B_T)^c$, the event that Peony loved someone else.

Conditional on the observed data,

$$
\begin{aligned}
P(A_T|\mathbf{n},\boldsymbol{\theta}) &= \mathbf{1}\{\theta_2 = -1\} \times (1-\theta_4)^{n_T} \times (1-\gamma)^{m_T} \\
P(A_T|\mathbf{n},\mathbf{X}=\mathbf{0},\mathbf{Z}=\mathbf{0},\boldsymbol{\theta}) &= \frac{P(A_T,\mathbf{X}=\mathbf{0},\mathbf{Z}=\mathbf{0}|\mathbf{n},\boldsymbol{\theta})}{P_{XZ}} \\
&= \frac{P(\mathbf{X}=\mathbf{0},\mathbf{Z}=\mathbf{0}|A_T) \times P(A_T|\mathbf{n},\boldsymbol{\theta})}{P_{XZ}} \\
&= \frac{1 \times \mathbf{1}\{\theta_2 = -1\} \times (1-\theta_4)^{n_T} \times (1-\gamma)^{m_T}}{P_{XZ}} \\
P(B_T|\mathbf{n},\mathbf{X}=\mathbf{0},\mathbf{Z}=\mathbf{0},\boldsymbol{\theta}) &= \frac{P(B_T,\mathbf{X}=\mathbf{0},\mathbf{Z}=\mathbf{0}|\mathbf{n},\boldsymbol{\theta})}{P_{XZ}} \\
&= \frac{1}{P_{XZ}}\Bigg[(1-\theta_3)^{m_T} \times \mathbf{1}\{\theta_2 = 1\} \\
&\quad + \sum_{j=1}^{m_T} \gamma(1-\gamma)^{j-1}(1-\theta_4)^{n(j)} \times (1-\theta_3)^{m_T-j+1} \times \mathbf{1}\{\theta_2 = -1\}\Bigg]
\end{aligned}
$$

Then,

$$
\begin{aligned}
&P(A_T|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0}) \\
&= \sum_{\theta_2=-1}^{N} \sum_{\theta_1=\max\{\theta_2,n_T\}}^{N} \int_0^1 \int_0^1 P(A_T|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0},\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0})\mathrm{d}\theta_3\mathrm{d}\theta_4 \\
&= \sum_{\theta_2=-1}^{N} \sum_{\theta_1=\max\{\theta_2,n_T\}}^{N} \int_0^1 \int_0^1 \frac{\mathbf{1}\{\theta_2 = -1\}(1-\theta_4)^{n_T}(1-\gamma)^{m_T}}{P_{XZ}} \times c_T(1+\theta_1)^{\mathbf{1}\{\theta_2=-1\}} D(\theta_1)(1-\theta_4)^{\beta-1} P_{XZ}\mathrm{d}\theta_3\mathrm{d}\theta_4 \\
&= \sum_{\theta_1=n_T}^{N} c_T(1+\theta_1)D(\theta_1) \times (1-\gamma)^{m_T} \int_0^1 \int_0^1 (1-\theta_4)^{n_T+\beta-1}\mathrm{d}\theta_3\mathrm{d}\theta_4 \\
&= \sum_{\theta_1=n_T}^{N} \frac{c_T(1+\theta_1)D(\theta_1)(1-\gamma)^{m_T}}{n_T+\beta}
\end{aligned}
$$

$$
\begin{aligned}
&P(B_T|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0})\\
&=\sum_{\theta_2=-1}^{N}\sum_{\theta_1=\max\{\theta_2,n_T\}}^{N}\int_0^1\int_0^1 P(B_t|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0},\boldsymbol{\theta})\times\pi(\boldsymbol{\theta}|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0})\mathrm{d}\theta_3\mathrm{d}\theta_4\\
&=\sum_{\theta_1=n_T}^{N} c_T D(\theta_1)\int_0^1\int_0^1(1-\theta_3)^{m_T}(1-\theta_4)^{\beta-1}\mathrm{d}\theta_3\mathrm{d}\theta_4\\
&\qquad+\sum_{\theta_1=n_T}^{N}\sum_{j=1}^{m_T} c_T\gamma(1-\gamma)^{j-1}(1+\theta_1)D(\theta_1)\int_0^1\int_0^1(1-\theta_4)^{\beta-1+n(j)}(1-\theta_3)^{m_T-j+1}\mathrm{d}\theta_3\mathrm{d}\theta_4\\
&=\sum_{\theta_1=n_T}^{N}\frac{c_T D(\theta_1)}{\beta(m_T+1)}+\sum_{\theta_1=n_T}^{N}\sum_{j=1}^{m_T}\frac{c_T\gamma(1-\gamma)^{j-1}(1+\theta_1)D(\theta_1)}{[\beta+n(j)](m_T-j+2)}
\end{aligned}
$$

And $P(C_T|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0})=1-P(A_T|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0})-P(B_T|\mathbf{X}=\mathbf{0},\mathbf{n},\mathbf{Z}=\mathbf{0})$. All of them are in summation form and thus can be obtained with simple programming.

2.3 Result and discussion

We computed the numerical result of our targets up to day $T=15$. In this part, we will discuss the posterior probability of Congwen being Peony's true love (event A_T) according to our previous setting.

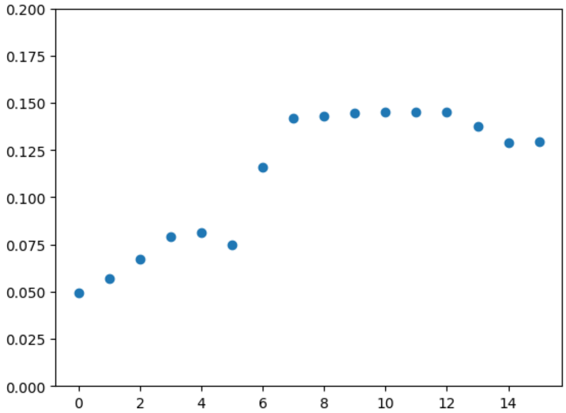


Figure 2: Plot of posterior probability of event A_T (x-axis is day T)

As Figure 2 shows, the probability increases from 0.05 to around 0.08 before day 4 but slightly drops on day 5, similar to the lecture's example. The initiative thought is that Congwen believes Peony will love him as she does not reply to people who send her letters. However, since more and more people send letters to Peony and Peony does not respond to these letters, Congwen may lose his confidence and consider that Peony may not love him and that she only focuses on study.

But unlike the lecture's example, when Congwen first sends the first love letter to Peony on Day 6, the posterior probability increases significantly from 0.075 to around 0.11. The reason is that there is a chance that Peony changes her mind and fall in love with Congwen once she receives his love letter. On Day 7, Congwen keeps sending the letter, and the probability increases again.

The probability remains stable in the second week until Congwen sends the third and the fourth letters. This time, unlike the previous one, the probability dips a bit. This reflects that if Peony loves Congwen, it is quite unlikely that she does not reply to his letter that many times. So it is more probable that she loves someone else.

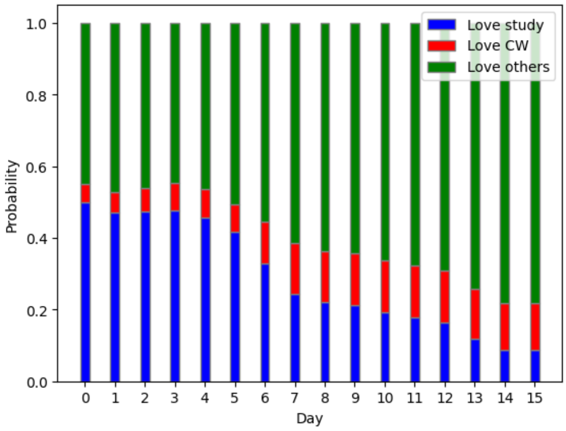


Figure 3: Plot of Probability for Event A, B, and C

We also take a look at the probabilities of the three target events (Figure 3). We can see the probability that Peony does not have true love (or love study) decreases with time and the number of love letters she receives. Simultaneously, the probability of her changing her mind to prefer Congwen or someone else gradually increases because of the letters she receives.

Compared to the old model in the lecture, our new model considers the possibility that a love letter can touch one's heart and make Peony fall in love. The result matches with the reality that if one sends one or two love letters, it can touch her heart, but if he keeps sending while she doesn't reply, it is more likely that she has loved someone else. However, Shen Congwen could always keep writing love letters, because the chance that Peony switches to loving him is always non-zeros.

3 Project 2: Rather die than compromise

3.1 Background story and settings

Yan Huizhu is a famous Kun opera actress who started her career in 1939. After she started her career life, she attempted suicide three times due to various reasons and finally died after her third suicide in 1966. During her career life, she did not try to kill her life when she encountered a bad mind each time. She made 7 compromises to her mind and continued her life.

The inter-arrival time of the i-th suicide is denoted as S_i, and the inter-arrival time of the i-th compromise is denoted as C_i. We assume an i.i.d. but nonparametric setting, $S_1, S_2, ... \overset{\text{iid}}{\sim}$ cdf F and $C_1, C_2, ... \overset{\text{iid}}{\sim}$ cdf G.

In this project, our target is to estimate the probability of no compromise between the first two suicides $\theta(F, G)$, assuming Huizhu has just begun a new career in a parallel universe, but using the same statistics model as the previous one. We will first obtain the maximum likelihood estimator of the distribution, i.e., $\tilde{F}$, $\tilde{G}$, and then use $\theta(\tilde{F}, \tilde{G})$ as our estimate. We finally conduct a Bootstrap sampling to estimate the performance of $\theta(\tilde{F}, \tilde{G})$, which are bias, mean squared error (MSE), standard deviation and median absolute deviation (MAD).

We consider two scenarios. The first scenario is that time now is supposed to be in 1966 when Huizhu had just had her 3rd suicide. The second scenario is that the time now is supposed to be in 1965. The key difference is that the first scenario fixes the times of suicide $m = 3$, while the second one fixes the time span from 1939 to 1965, and thus m is a random variable M as N.

Data obtained:
Counting from the year 1939 to 1966, We observe $m = 3$ suicides $N = 7$ compromises. The interval of suicide S_i and compromise C_i is shown in table 2.

Denote $\mathbf{S} = [S_1, S_2, S_3]^T$ and $\mathbf{C} = [C_1, \cdots, C_N]^T$. Note that in scenario 2, we do not observe S_3, and $m = 2$.

	Interval	**Year**
S_1	6	1945
S_2	10	1955
S_3	11	1966
C_1	10	1949
C_2	5	1954
C_3	1	1955
C_4	1	1956
C_5	1	1957
C_6	0	1957
C_7	7	1964

Table 2: Values of Inter-arrival Time and Corresponding Event Years

3.2 Maximum likelihood estimate of probability distribution

3.2.1 Scenario 1

We can obtain the likelihood function for scenario 1,

$$l_1(F, G) = \prod_{i=1}^{m} P_F(X = S_i) \times \prod_{i=1}^{N} P_G(Y = C_i) \times P_G\left(Y > \sum_{i=1}^{m} S_i - \sum_{i=1}^{N} C_i | \mathbf{S}, \mathbf{C}\right)$$

Denote

1. $P_F(i) = P_F(X = S_i)\ \forall 1 \le i \le m$
2. $P_G(i) = P_F(X = C_i)\ \forall 1 \le i \le N$
3. $\Delta_1 = \sum_{i=1}^{m} S_i - \sum_{i=1}^{N} C_i$, and $v := P_G(Y > \Delta_1)$

We construct the program to obtain the maximum likelihood estimate of F and G,

$$\begin{aligned}
\text{Maximize } & v\prod_{i=1}^{m} p_F(i)\prod_{i=1}^{N} p_G(i) \\
\text{Subject to } & \sum_{i=1}^{m} p_F(i) \leq 1 \\
& 1-v \geq \sum_{i=1}^{N} p_G(i)\mathbf{1}\{C_i \leq \Delta_1\} \\
& v \geq \sum_{i=1}^{N} p_G(i)\mathbf{1}\{C_i > \Delta_1\} \\
& p_F(i) \geq 0 \ \forall 1 \leq i \leq m \\
& p_G(i) \geq 0 \ \forall 1 \leq i \leq N, \ 0 \leq v \leq 1
\end{aligned}$$

First, observe that at the optimum, the first constraint should be tight, and the method of Lagrange multiplier gives that $p_F(i) = \frac{1}{m}\ \forall 1 \leq i \leq m$. Thus, the maximum likelihood estimate $\tilde{F}$ is a uniform distribution among $\{S_i\}_{i=1}^{m}$. Next, we optimize $v\prod_{i=1}^{N} p_G(i)$. Let N_1 be the number of observations in $\mathbf{C}$ that $C_i \leq \Delta_1$. Two cases should be considered separately.

1. Case 1: when $N_1 = N$, which means there is no observed C_i greater than Δ_1. The program becomes

$$\begin{aligned}
\text{Maximize } & v\prod_{i=1}^{N} p_G(i) \\
\text{Subject to } & 1-v \geq \sum_{i=1}^{N} p_G(i) \text{ and } p_G(i) \geq 0 \ \forall 1 \leq i \leq N, \ 0 \leq v \leq 1
\end{aligned}$$

   By Lagrange multiplier, the optimum is attained at $v = p_G(i) = \frac{1}{N+1}\ \forall 1 \leq i \leq N$. Therefore, the maximum likelihood estimate $\tilde{G}$ should satisfy $P_{\tilde{G}}(Y > \Delta_1) = p_{\tilde{G}}(C_i) = \frac{1}{N+1}\ \forall 1 \leq i \leq N$. But we have no information for the part greater than Δ_1. We test on two cases, $p_{\tilde{G}}(\infty) = \frac{1}{N+1}$, which means the incident (suicide/compromise) will never happen thereafter (referred as infinity setting in the following context) and $p_{\tilde{G}}(\Delta_1 + 1) = \frac{1}{N+1}$ (referred as $\Delta + 1$ setting in the following context)

2. Case 2: when $N_1 < N$, the second and the third constraint is necessarily tight at the optimum, and using Lagrange multiplier, $p_G(i)$ takes identical value $\frac{1-v}{N_1}$ for i such that $C_i \leq \Delta_1$, and

takes $\frac{v}{N-N_1}$ for i such that $C_i > \Delta_1$. The program is simplified into

$$\text{Maximize } v\left(\frac{1-v}{N_1}\right)^{N_1}\left(\frac{v}{N-N_1}\right)^{N-N_1}$$
$$\text{Subject to } 0 \le v \le 1$$

which gives the optimal point $v = \frac{N+1-N_1}{N+1}$ and then $p_G(i) = \frac{1}{N+1}$ for i such that $C_i \le \Delta_1$, and $p_G(i) = \frac{1}{N+1}(1+\frac{1}{N-N_1})$ for i such that $C_i > \Delta_1$. Thus we obtain the exact $\tilde{G}$.

Given the data

$S_1 = 6, S_2 = 10, S_3 = 11, N = 7, N_1 = 4, C_1 = 10, C_2 = 5, C_3 = C_4 = C_5 = 1, C_6 = 0, C_7 = 7, \Delta_1 = 2$

which corresponds to case 2, we have

$$p_{\tilde{F}}(X=6) = p_{\tilde{F}}(X=10) = p_{\tilde{F}}(X=11) = \frac{1}{3}$$
$$p_{\tilde{G}}(X=0) = \frac{1}{8}, p_{\tilde{G}}(X=1) = \frac{3}{8}, p_{\tilde{G}}(X=5) = p_{\tilde{G}}(X=7) = p_{\tilde{G}}(X=10) = \frac{1}{6}$$

3.2.2 Scenario 2

In scenario 2, m becomes random, now denoted as M, such that the year 1965 lies between the M-th and the $M+1$-th suicides. Denote $T = 1965 - 1939 = 26$ be the time span from the beginning. The likelihood function of our observation {M,**S**,N,**C**} is

$$l_2(F,G) = \prod_{i=1}^{M} P_F(X=S_i) \times P_F\left(X > T - \sum_{i=1}^{M} S_i|\mathbf{S}\right) \times \prod_{i=1}^{N} P_G(Y=C_i) \times P_G\left(Y > T - \sum_{i=1}^{N} C_i|\mathbf{C}\right)$$

Since $\{\mathbf{S}, M\}$ and $\{\mathbf{C}, N\}$ are independent, we can optimize $\prod_{i=1}^{M} P_F(X=S_i)P_F\left(X > T - \sum_{i=1}^{M} S_i\right)$ for $\tilde{F}$ and $\prod_{i=1}^{N} P_G(Y=C_i) \times P_G\left(Y > T - \sum_{i=1}^{N} C_i\right)$ for $\tilde{G}$ separately in a similar way.

We do the first half for $\tilde{F}$, while the second half for $\tilde{G}$ is the same. Denote $\Delta_2 := T - \sum_{j=1}^{M} S_j$ and $v' := T - P_F(X > \Delta_2)$. We can find that optimizing $v\prod_{i=1}^{M} P_F(X=S_i)$ is exactly what we have done in scenario 1, which considers these two cases on M_1, the number of observations in **S** that $S_i \le \Delta_2$.

1. Case 1: $M_1 = M$. The maximum likelihood $\tilde{F}$ is $v' = p_{\tilde{F}}(X=S_i) = \frac{1}{M+1}$ $\forall 1 \le i \le M$. And we also test both the infinity setting ($p_{\tilde{F}}(X=\infty) = \frac{1}{M+1}$) and the $\Delta+1$ setting ($p_{\tilde{F}}(X=\Delta_2+1) = \frac{1}{M+1}$).

2. Case 2: $M_1 < M$. The maximum likelihood $\tilde{F}$ is $p_{\tilde{F}}(S_i) = \frac{1}{M+1}$ for i such that $S_i \leq \Delta_2$ and $p_{\tilde{F}}(S_i) = \frac{1}{M+1}(1 + \frac{1}{M-M_1})$ for i such that $S_i > \Delta_2$.

Similar for $\tilde{G}$ with the corresponding $\Delta_3 := T - \sum_{i=1}^{N} C_i$ and N_1',

$$p_{\tilde{G}}(y) = \begin{cases} \frac{1}{N+1} & , \quad x \in \mathbf{C} \text{ and } x \leq \delta_3 \\ \frac{1}{N+1}(1 + \frac{1}{N-N_1'}) & , \quad x \in \mathbf{C} \text{ and } x > \Delta_3 \\ \frac{1}{N+1} & , \quad x = \Delta_3 + 1 \text{ or } \infty, \text{ and } N_1' = N \end{cases}$$

(We abuse the notation a bit to use $\mathbf{C}$ to denote the (multi-)set of its entries.) Note that the choice of $\Delta_3 + 1$ or ∞ depends on whether we adopt the infinity setting or the $\Delta + 1$ setting.
From the data

$$M = 2, M_1 = 2, S_1 = 6, S_2 = 10, \Delta_2 = 10$$

which corresponds to case 1, we have

$$p_{\tilde{F}}(X = 6) = p_{\tilde{F}}(X = 10)$$

and the remaining $\frac{1}{3}$ probability is assigned to either 11 or ∞ depend on the setting.
From the data

$$N = 7, N_1' = 4, C_1 = 10, C_2 = 5, C_3 = C_4 = C_5 = 1, C_6 = 0, C_7 = 7, \Delta_3 = 1$$

which corresponds to case 2, we obtain

$$p_{\tilde{G}}(X = 0) = \frac{1}{8}, p_{\tilde{G}}(X = 1) = \frac{3}{8}, p_{\tilde{G}}(X = 5) = p_{\tilde{G}}(X = 7) = p_{\tilde{G}}(X = 10) = \frac{1}{6}$$

3.3 Estimation of target

3.3.1 Estimation of θ

We estimate $\theta(F, G)$ with

$$\theta(\tilde{F}, \tilde{G}) = P\left[\exists k, \sum_{i=1}^{k} C_i' \leq S_1' \text{ and } \sum_{i=1}^{k+1} C_i' \geq S_2'\right]$$

where $S_1', S_2' \overset{i.i.d}{\sim} \tilde{F}$ and $C_1', C_2', \cdots \overset{i.i.d}{\sim} \tilde{G}$. Note that in the infinity setting, it is possible for either S' or C' to take infinity. From the natural meaning, we suppose that the target probability excludes the cases S_1' or S_2' taking ∞, where the second suicide does not exist, while it includes the case C_{k+1}'

taking ∞, where there would be no longer a compromise.
In our sample, distribution $\tilde{G}$ can take value 0 with positive probability. Given any S_1 and S_2, it is possible that the sequence $\{C'_i\}$ takes up to infinite step k before $\sum_{i=1}^{k+1} C'_i > S'_1$. Thus, the target is a sum of infinity value. We employ two methods to arrive at the target, which is an adaptive summing method with accuracy control and a Monte Carlo simulation.

The first method starts with the following recursive formula on the probability

$$f(t, y) := P_{\tilde{G}} \left(\sum_{i=1}^{t} C_i = y \right)$$

we have the boundary case $f(0, 0) = 1$, $f(0, y) = 0$ $\forall y \geq 1$ and

$$\forall t \geq 0, y \geq 0,\ f(T, y) = \sum_{z=0}^{y} f(t-1, z) \times P_{\tilde{G}}(C_t = y - z)$$

Remark: In practice, enumerating the value that $\tilde{G}$ can take with positive probability is more efficient than enumerating from 0 to y.
Our target can be formularized as

$$\begin{aligned} \theta(\tilde{F}, \tilde{G}) &= \sum_{S_1} \sum_{S_2} P_{\tilde{F}}(S_1) P_{\tilde{F}}(S_2) P_{\tilde{G}} \left[\text{series of } C_i \text{ have no value in } (S_1, S_1 + S_2)\right] \\ &= \sum_{S_1} \sum_{S_2} P_{\tilde{F}}(S_1) P_{\tilde{F}}(S_2) \sum_{C} P_{\tilde{G}}(C) \sum_{y=\max\{0, S_1+S_2-C\}}^{S_1} \sum_{t=0}^{\infty} f(t, y) \end{aligned}$$

where the summing of S_1, S_2 and C is in the range of the values with positive probability in the corresponding distribution. Note that the choice of S_1, S_2 should exclude ∞, but C should not. The summing of t up to infinity is conducted in an adaptive way, which will stop whenever $f(t, y)$ for all y are less than $\epsilon = 10^{-6}$. In our experiment, it can be achieved within 20 iterations of t. This method can obtain the actual answer with a controllable error bound (though the error bound for the target $\theta(\tilde{F}, \tilde{G})$ of this method is not explicitly computed, we believe it is guaranteed with ϵ, since we have tried different ϵ ranging from 10^{-4} to 10^{-7} and the result is not significantly different).

For the Monte Carlo simulation, We generate $n = 200$ sample of S'_1, S'_2 and $\{C_i\}$. In each sample, $\{C'_i\}$ is obtained by randomly choosing a value in $\tilde{G}$ recursively until $\sum_{i=1}^{k+1} C'_i > S'_1$, i.e., we can confirm whether $\sum_{i=1}^{k+1} C'_i$ lies in (S'_1, S'_2). $\theta(\tilde{F}, \tilde{G})$ is estimated to be the proportion of sample that $\sum_{i=1}^{k} C'_i \leq S'_1$ and $\sum_{i=1}^{k+1} C'_i \geq S'_2$ for some k.

In our experiment, the variation of the Monte Carlo estimate can be up to 10% of the target value, which shows that $n = 200$ is not enough. However, we can hardly further increase n due to prolonged running time in the bootstrap estimation.

3.3.2 Bootstrap estimation

In this section, we will present the procedure of bootstrap estimation of the performance of our estimator $\theta(\tilde{F}, \tilde{G})$. First, we will construct bootstrap cdf. of estimation error $\theta(\tilde{F}, \tilde{G}) - \theta(F, G)$. The bootstrap analogue is $\theta(\tilde{F}^*, \tilde{G}^*) - \theta(\tilde{F}, \tilde{G})$, where $\tilde{F}^*, \tilde{G}^*$ is the maximum likelihood estimate of the boostrap sample $\mathbf{S}^*$ and $\mathbf{C}^*$.

Bias, MSE, Standard deviation, and MAD of $\theta(\tilde{F}, \tilde{G})$ can be approached with $B = 1000$ bootstrap samples. And the cdf. of $\theta(\tilde{F}, \tilde{G}) - \theta(F, G)$ can also be approximate by the cdf. of $\theta(\tilde{F}^*, \tilde{G}^*) - \theta(\tilde{F}, \tilde{G})$

$$\text{Bias} \approx B^{-1} \sum_{b=1}^{B} (\tilde{\theta}^{*b} - \theta(\tilde{F}, \tilde{G}))$$

$$\text{MSE} \approx B^{-1} \sum_{b=1}^{B} (\tilde{\theta}^{*b} - \theta(\tilde{F}, \tilde{G}))^2$$

$$\text{Standard deviation} \approx \sqrt{\text{MSE} - \text{Bias}^2}$$

$$\text{MAD} \approx B^{-1} \sum_{b=1}^{B} |\widetilde{\theta^{*b}} - \theta(\tilde{F}, \tilde{G})|$$

$$\text{cdf}(t) \approx B^{-1} \sum_{b=1}^{B} \mathbf{1}\{\tilde{\theta}^{*b} - \theta(\tilde{F}, \tilde{G}) \leq t\}$$

We need to discuss some technical details of generating the bootstrap samples. In scenario 1, we first sample S_1^*, S_2^*, S_3^* from $\tilde{F}$ with replacement, then generate a stream of $C_1^*, C_2^*, \ldots$ i.i.d. from $\tilde{G}$ until $\sum_{i=1}^{N^*} C_i^* \leq \sum_{j=1}^{3} S_j^* < \sum_{i=1}^{N^*+1} C_i^*$. In scenario 2, we generate both sequence $S_1^*, S_2^*, \ldots$ and $C_1^*, C_2^*, \ldots$ from $\tilde{F}$ and $\tilde{G}$ respectively until $\sum_{j=1}^{M^*+1} S_j^* > T$ and $\sum_{i=1}^{N^*+1} C_i^* > T$ ($T = 26$). The corresponding M^* and N^* are also a part of the sample. After getting the sample $(M^*, \mathbf{S}^*, N^*, \mathbf{C}^*)$, $\tilde{F}^*$, $\tilde{G}^*$ and $\theta(\tilde{F}^*, \tilde{G}^*)$ are estimated in the same way as the previous part. Two settings and four methods are considered.

3.4 Result and discussion

In this final section, we will present our numerical result of both scenarios, each with 4 methods, which are infinity setting+adaptive summing, infinity setting+Monte Carlo, $\Delta+1$ setting+adaptive summing, $\Delta+1$ setting+Monte Carlo.

3.4.1 Scenario 1 Result

In Scenario 1, the time now is in 1966 when Yan Huizhu had just had her 3rd suicide. We adapt the four methods mentioned above and obtain the results of nonparametric mle, $\theta(\tilde{F},\tilde{G})$ respectively.

Method 1: Infinity setting, Monte Carlo method
Method 2: Infinity setting, Adaptive summing method
Method 3: $(\Delta+1)$ setting, Monte Carlo method
Method 4: $(\Delta+1)$ setting, Adaptive summing method

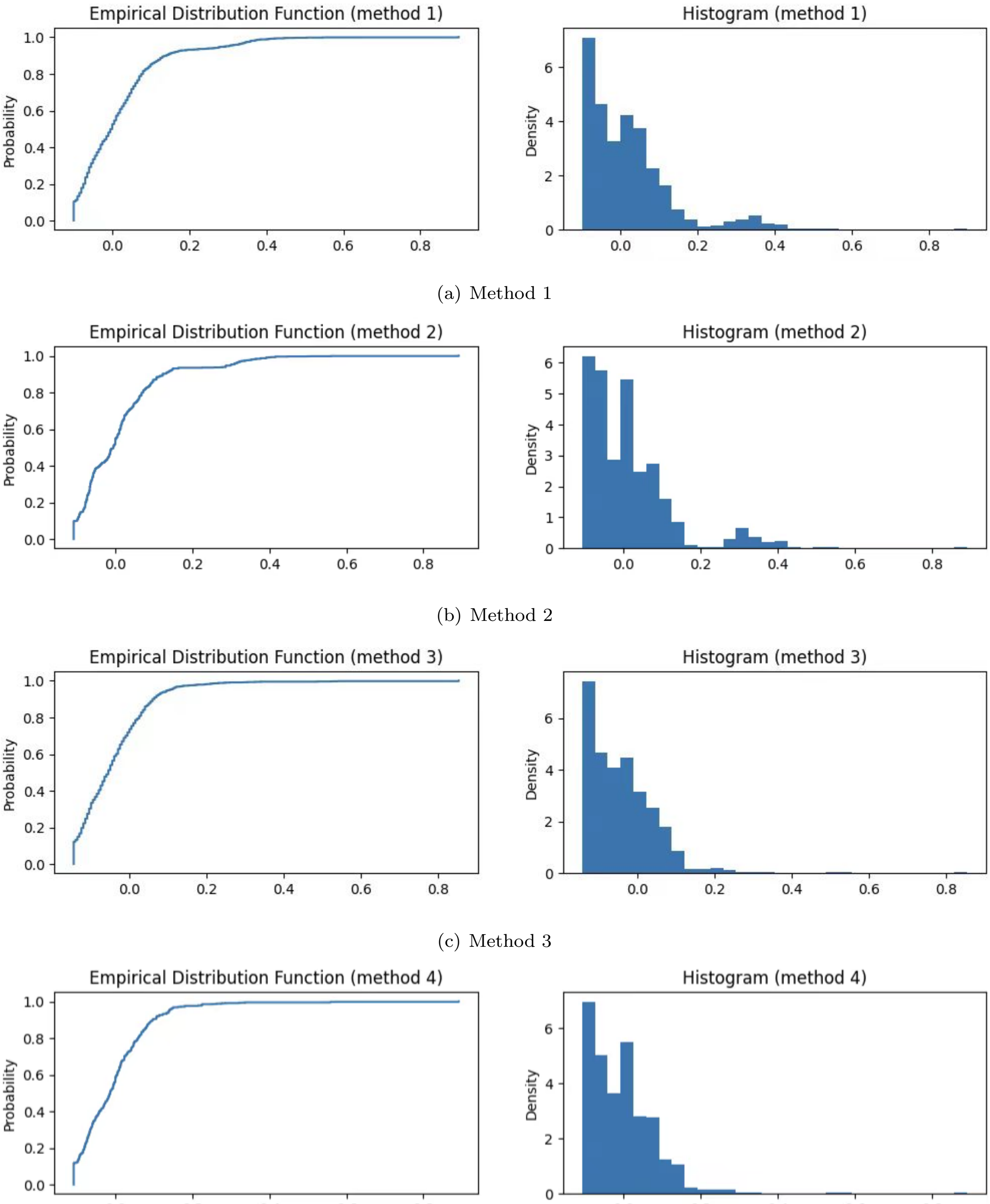
The results of each method are shown below. Figure 4 shows the cdf and histograms.

Table 3: Table of Estimates Obtained by Each Method in Scenario 1

	$\theta(\tilde{F},\tilde{G})$	Bias	MSE	SD	MAD
Method 1	0.100	0.0189	0.014	0.115	0.081
Method 2	0.108	0.011	0.0131	0.114	0.079
Method 3	0.145	-0.042	0.011	0.096	0.082
Method 4	0.108	-0.004	0.009	0.095	0.068

From the table, we can observe that the results we obtain through four methods are generally consistent. Compared to the estimated value (around 0.1), the bootstrap estimation of bias is small, but the MSE, SD, and MAD are relatively big (around the size of the estimate), which indicates that the estimator is approximately unbiased, but has a large variance, mainly due to the small data size.

We notice that when we use the adaptive summing method, the estimations obtained are equal under the infinity setting and $\Delta+1$ setting, which should be so because our data do not have the condition to differentiate them. However, using Monte Carlo methods, the estimates have noticeable differences because of the nature of randomness and inadequate sample size.



(a) Method 1

(b) Method 2

(c) Method 3

(d) Method 4

Figure 4: Plots of Empirical Distribution Function and Histogram in Scenario 1

3.4.2 Scenario 2 Result

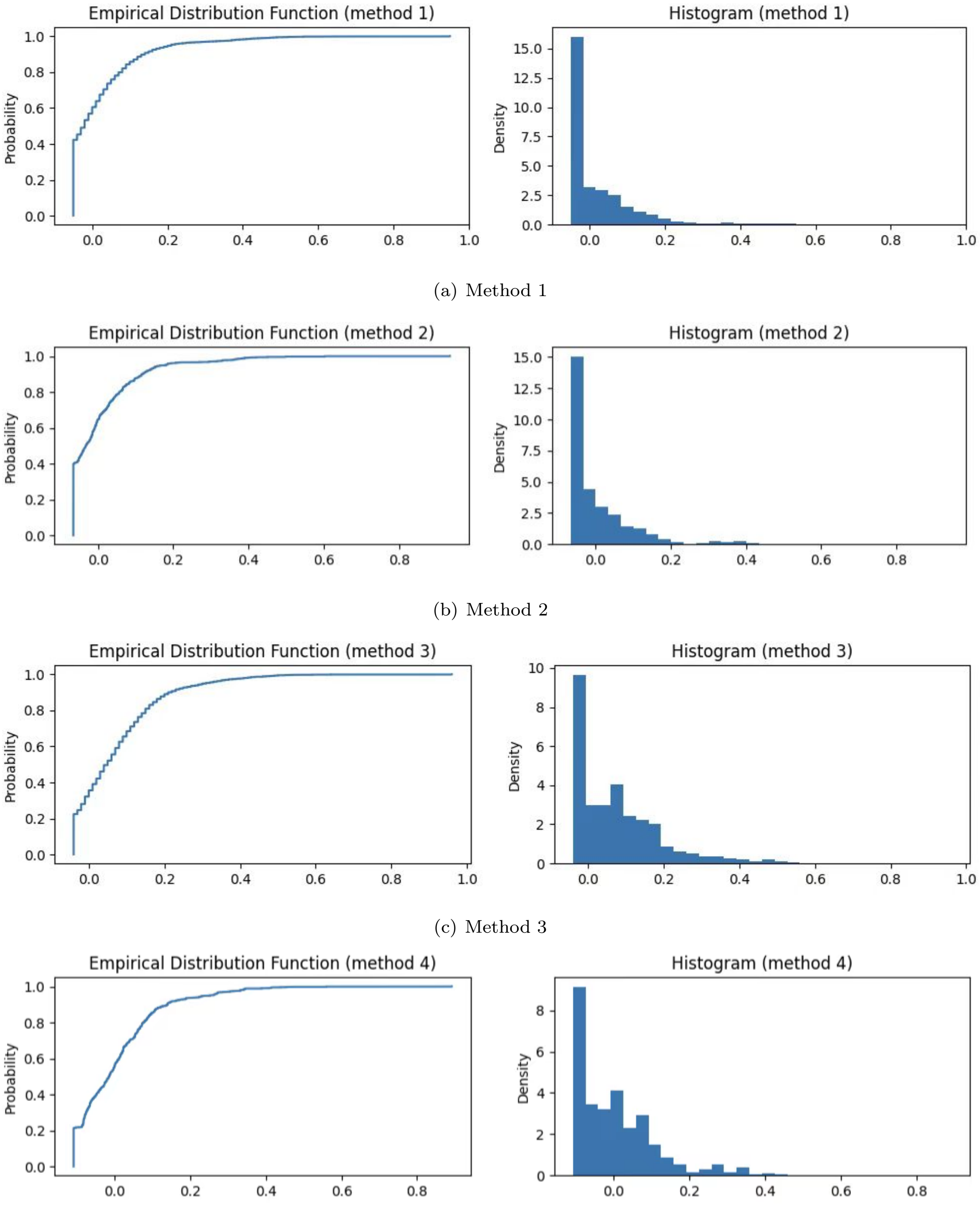
The results of the four methods in scenario 2 are shown in Table 4 and Figure 5.

Table 4: Table of Estimates Obtained by Each Method in Scenario 2

	$\theta(\tilde{F}, \tilde{G})$	Bias	MSE	SD	MAD
Method 1	0.050	0.019	0.011	0.103	0.068
Method 2	0.065	0.003	0.010	0.099	0.069
Method 3	0.040	0.073	0.019	0.117	0.094
Method 4	0.108	0.005	0.013	0.113	0.084

Still, only the results of adaptive summing method (methods 2 and 4) are informative because of its accuracy. We can observe that under $\Delta + 1$ setting, the estimate is the same as scenario 1, while the one in the infinity setting is smaller. This makes sense because the second setting allows the second suicide does not exist, which does not contribute to our target probability.

Also, the same as scenario 1, the estimators are approximately unbiased but have a relatively large variance.



(a) Method 1

(b) Method 2

(c) Method 3

(d) Method 4

Figure 5: Plots of Empirical Distribution Function and Histogram in Scenario 2