

12-2023

Parameter Estimation for Patient Enrollment in Clinical Trials

Junyan Liu
William & Mary

Follow this and additional works at: <https://scholarworks.wm.edu/honorstheses>



Part of the [Applied Mathematics Commons](#), [Applied Statistics Commons](#), [Data Science Commons](#), [Operational Research Commons](#), and the [Survival Analysis Commons](#)

Recommended Citation

Liu, Junyan, "Parameter Estimation for Patient Enrollment in Clinical Trials" (2023). *Undergraduate Honors Theses*. William & Mary. Paper 2066.

<https://scholarworks.wm.edu/honorstheses/2066>

This Honors Thesis -- Open Access is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.

Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from the College of William and Mary, I hereby grant to the College of William and Mary and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter known, including display online. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

A handwritten signature in black ink, featuring stylized Chinese characters and the English name 'Junyan'.

November 19, 2023

Junyan Liu

Date

Parameter Estimation for Patient Enrollment in Clinical Trials

A thesis presented in Candidacy for Departmental Honors in
Computational and Applied Mathematics and Statistics

from

The College of William and Mary in Virginia

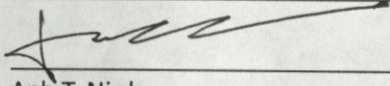
By

Junyan Liu

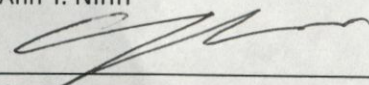
Dec 08, 2023

Accepted for

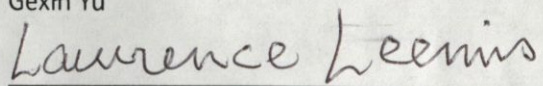
Honors



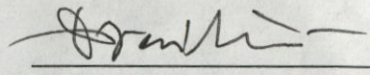
Anh T. Ninh



Gexin Yu



Lawrence M. Leemis



Daniel Vasiliu

Parameter Estimation for Patient Enrollment in Clinical Trials

By

Junyan Liu

Advisor: Anh Ninh

An abstract of
A thesis submitted to the Faculty of the
College of William and Mary
in partial fulfillment of the requirements for the degree of
Bachelor of Science
in Department of Mathematics
2023

Abstract

Parameter Estimation for Patient Enrollment in Clinical Trials

By Junyan Liu

In this paper, we study the Poisson-gamma model for recruitment time in clinical trials. We proved several properties of this model that match our intuitions from a reliability perspective, did simulations on this model, and used different optimization methods to estimate the parameters. Although the behaviors of the optimization methods were unfavorable and unstable, we identified certain conditions and provided potential explanations for this phenomenon and further insights into the Poisson-gamma model.

Parameter Estimation for Patient Enrollment in Clinical Trials

By

Junyan Liu

Advisor: Anh Ninh

A thesis submitted to the Faculty of the
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements for the degree of
Bachelor of Science
in Department of Mathematics
2023

Acknowledgments

I would like to express my deepest gratitude to my thesis advisor, Professor Anh Ninh, for his invaluable guidance, unwavering support, and insightful feedback throughout the research process. His expertise and encouragement have been instrumental in shaping this work.

I am also thankful to the members of my thesis committee, Professor Larry Leemis, Professor Gexin Yu, and Professor Daniel Vasiliu, for their time and valuable suggestions, which greatly enhanced the quality of this thesis.

I extend my heartfelt appreciation to my family and friends, especially my mother, Chen Xingping; my grandmother, Zeng Xiaoqiong, for their endless love, encouragement, and understanding. Their patience and belief in my abilities sustained me through the challenging phases of this research journey.

I also acknowledge the support provided by the College of William and Mary for its resources that made it possible for me to dedicate my time and effort to this project.

Finally, this thesis was only possible with the collective support, encouragement, and understanding of all these individuals and intuitions. Thank you for participating in this significant milestone in my academic journey.

Junyan Liu

College of William and Mary

November 19, 2023

Contents

1	Introduction	1
2	Recruitment Modeling	3
2.1	Notation and Definitions	3
2.2	Mathematical Properties	6
3	Estimation Problems	12
3.1	Simulated Study	12
3.2	Estimation methods	14
3.3	Optimization Problem	17
4	Numerical Studies	19
4.1	Performance of Optimization Algorithms	19
4.2	Robustness of Algorithms	23
5	Conclusion and Future Research	26
A	Appendix	28
	Bibliography	29

List of Figures

2.1	The survivor function of the recruitment time in a multi-site trial with $\alpha = 100, \beta = 50, n = 100, N = 5$ (black) and $\alpha = 8, \beta = 4, n = 100, N = 5$ (red).	10
2.2	On the left is the survivor function of the recruitment time in a multi-site trial with $\alpha = 100, \beta = 50, n = 130, N = 6$ (black) and $\alpha = 80, \beta = 40, n = 100, N = 5$ (red). On the right is the survivor function of the recruitment time in a multi-site trial with $\alpha = 100, \beta = 50, n = 130, N = 6$ (black) and $\alpha = 8, \beta = 4, n = 100, N = 5$ (red).	10
3.1	The PearsonVI distribution of T and the histogram of the simulated recruitment time with $\alpha = 10, \beta = 10, n = 200, N = 4$	13
3.2	Simulations with $\alpha = 4, \beta = 20, N = 20, T = 30$. Fit into a gamma distribution and obtain $\hat{\alpha} = 3.883$ and $\hat{\beta} = 17.360$	14
3.3	Simulations with $\alpha = 4, \beta = 20, N = 20, T = 30$ and using optimization to obtain fitted parameters with $\hat{\alpha} = 9.647$ and $\hat{\beta} = 43.122$	17
4.1	Simulations with $\alpha = 6, \beta = 6, N = 50, n = 200$, starting with initial parameters $\alpha = \beta = 2$ on the top and $\alpha = \beta = 10$ on the bottom. The numbers behind each name of optimization are $\hat{\alpha}$ and $\hat{\beta}$	20
4.2	Contour map of the objective function with α and β from 1 to 10 (simulations with $\alpha = 6, \beta = 6, N = 50, n = 200$).	22

- 4.3 Distributions of $\hat{\alpha}$ and $\hat{\beta}$ with simulations from $\alpha = 6$, $\beta = 6$, $n = 200$,
 $N = 50, 20, 5$ on the first, second and third row respectively. 24
- 4.4 Histogram of $\hat{\alpha}/\hat{\beta}$ with simulations from $\alpha = 6$, $\beta = 6$, $N = 50$, $n = 200$). 25

Chapter 1

Introduction

Clinical trials are essential research studies to evaluate the safety, efficacy, and effectiveness of new medical treatments, interventions, drugs, or devices in humans. A typical clinical trial usually takes certain phases with different scientific purposes, and this paper is trying to model patient recruitment time.

From time to time, when it comes to clinical trials, researchers and practitioners want to be as efficient as possible because it can not only save time but also relieve patients' pain and even save lives for any second saved, like releasing an effective vaccine during a pandemic. Swift recruitment leads to faster completion of trials, making new treatments available sooner to patients in need. From a scientific perspective, a precise estimation of recruitment time will make the process smoother, saving plenty of resources (especially time). Usually, the recruitment time problem seems unpredictable, so people have to wait for the recruitment to complete and plan the following phases of clinical after completion. Still, with a proper model to estimate recruitment time, we can significantly enhance our efficiency and allocate resources to other procedures.

Recruitment time is one of the crucial factors in clinical trials, so we need to consider enough factors for a more complete and precise prediction. In previous

studies, people have come up with different models for clinical trials. For instance, people first proposed the unconditional model[5], which estimates the time by dividing the acquired sample size by the number of recruited patients across all centers in one month. Then, it was improved to a conditional model, which allows the expected recruitment in any given month to vary, depending on other conditions in the trial like the number of recruitment centers[3]. However, these models are deterministic, so the patients arrive at the centers at certain rates but not randomly. Thus, for example, in [11], the author proposed a model where patients arrive randomly in different centers in Poisson processes with a fixed and identical recruitment rate. We use the revised version of this model, which assumes that the rates follow a gamma distribution because there are various uncertainties in different recruitment centers, so keeping the rate identical will not be useful given these large discrepancies across centers. Moreover, the gamma distribution can be improved using Bayesian analysis, so new evidence of data can help us improve our gamma model to have a better fit.

In this paper, we examined the Poisson-Gamma model proposed in the paper [2] in theoretical analysis, did simulations on the model, and showed several numerical results that are related and useful. Meanwhile, when trying to estimate and predict future results, we will reveal the instability in practice, identify the problems that occurred in the optimization process of this model, and point out potential ways to solve them.

Chapter 2

Recruitment Modeling

2.1 Notation and Definitions

We consider a Poisson-gamma recruitment model, where a clinical trial recruits patients at N clinical centers with independent Poisson processes, and the mean rates of these Poisson processes across different centers are from a gamma distribution. Denote by λ_i the recruitment rate at center i , and the recruitment is stopped when the total number of recruited patients reaches n , the desired number of recruited patients. In our model, all centers start recruitment simultaneously, so t is the time after the centers start recruiting.

Definition 1 (Poisson process). According to [7], the Poisson process is a stochastic process with $N(t)$, the counts of recruitment, with parameter $\lambda > 0$, that have the following properties :

- $N(0) = 0$.
- The process has independent increments.
- The number of events that occur in a fixed interval of time t follows a Poisson distribution with mean λt .

- $\lambda > 0$ is called the mean rate of a Poisson process.

Definition 2 (Poisson distribution). The Poisson distribution is a probability distribution that represents the number of events that occur within a fixed interval of time, with the probability mass function of the Poisson distribution given by the formula:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!},$$

where:

- $P(X = x)$ is the probability of observing x events in the interval.
- λ is the average number of events that occur in the given interval, which is the mean.
- x is the actual number of events $(0, 1, 2, \dots)$.

Definition 3 (Gamma distribution). The gamma distribution is a continuous probability distribution with the probability density function given by:

$$f(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)},$$

with $x > 0$, shape parameter $\alpha > 0$, and rate parameter $\beta > 0$. $\Gamma(\alpha)$ is the gamma function, defined as $\int_0^\infty t^{\alpha-1} e^{-t} dt$.

Intuitively, researchers want to compare recruitment time between different trials to see which is better. Hence, stochastic dominance is introduced for deciding the preference.

Definition 4 (First-order stochastic dominance). The random variable A has (first-order) stochastic dominance over random variable B if, for any x as an outcome, A gives at least as high a probability of getting an outcome at least x as B does.

In notation, $P[A \geq x] \geq P[B \geq x]$ for all x , and we denote A having stochastic dominance over B as $A \succeq_{FSD} B$. [10]

Between two recruitment T_1 and T_2 , if T_1 has stochastic dominance over T_2 , T_2 is preferred because it has a lower probability of having a more prolonged or equal recruitment time than t for any $t > 0$. To determine stochastic dominance, we can use the survival function.

Definition 5 (Survival function). For a random variable T on time t , the survival function $S(t)$ is defined [7] as:

$$S(t) = P(T \geq t), t > 0$$

In other words, the survival function $S(t)$ gives the probability that the random variable T takes a value greater than or equal to t .

Between two trials respectively with recruitment time T_1 and T_2 with $S_1(t)$ and $S_2(t)$ as their survival functions, we have the following[1]:

$$S_1(t) \geq S_2(t) \implies T_1 \succeq_{FSD} T_2$$

Returning to our problem, we may denote λ_i as the mean rate of center i in its Poisson process generated from a Gamma(α, β) distribution with shape parameter α and rate parameter β . In other words, independent Poisson processes are happening in different centers with the number of recruited patients $N_i(t)$ in the center i following the probability mass function:

$$P(N_i(t) = k) = \frac{(\lambda_i t)^k e^{-\lambda_i t}}{k!}, k = 0, 1, 2, \dots$$

Thus, to predict the total recruitment time to reach n patients from N centers whose means of Poisson processes are from Gamma(α, β), we want to study the

distribution of T , the aggregate recruitment time, when $N(t) = \sum_{i=1}^N N_i(t) = n$.

According to theoretical results from [2], the recruitment time T that these centers recruit an aggregated number of n patients follows a Pearson Type VI distribution, whose probability density function is given by:

$$f(t; \alpha, \beta, n, N) = \frac{1}{\mathcal{B}(n, \alpha N)} \frac{t^{n-1} \beta^{\alpha N}}{(t + \beta)^{n + \alpha N}},$$

where $t > 0$, $\mathcal{B}(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$ is the beta function, n is the number of patients of recruitment, and N is the number of sites.

2.2 Mathematical Properties

Intuitively, for recruitment time $T(n, N, \alpha, \beta)$, the mean is decreasing in α and increasing in β , as a lower α or higher β results in smaller means in the original gamma distribution, so λ_i generated from it will be lower, so it requires more time for the centers to recruit enough people. Moreover, the recruitment time with a lower α (or higher β) has stochastic dominance over the recruitment time with a higher α (or a lower β), given n and N fixed. On the other hand, with fixed α and β , the recruitment time with a lower N (or higher n) has stochastic dominance over the recruitment time with a higher N (or a lower n), and this is intuitive because more patients and fewer sites require more time for the completion of recruiting patients.

Now we study certain behaviors of tuning parameters on the distribution to match the intuition with proof of stochastic properties, with all parameters being positive integers.

Proposition 1. Assume that there are two clinical trials A and B , where trial A has N_1 centers and trial B has N_2 centers ($N_1 < N_2$), and the other parameters are the same. The recruitment time of trial A has stochastic dominance over trial B . Denote

the survival functions of trial A and B by $S_A(t)$ and $S_B(t)$ respectively.

Proof. We use the survivor function of the trials:

$$S(t) = \frac{\beta^{\alpha N}}{\mathcal{B}(n, \alpha N)} \int_t^\infty \frac{x^{n-1}}{(x + \beta)^{n+\alpha N}} dx = I_{\frac{\beta}{t+\beta}}(\alpha N, n)$$

where incomplete beta function is defined by $\mathcal{B}_t(a, b) = \int_0^t x^{a-1}(1-x)^{b-1} dx$ and $I_t(a, b) = \mathcal{B}_t(a, b)/\mathcal{B}(a, b)$. It is also well-known that [9]:

$$I_{\frac{\beta}{t+\beta}}(\alpha N, n) = I_{\frac{\beta}{t+\beta}}(\alpha N + 1, n) + \frac{(\frac{\beta}{t+\beta})^{\alpha N} (\frac{t}{t+\beta})^n}{\alpha N \cdot \mathcal{B}(\alpha N, n)}$$

For any fixed $t > 0$, all other parameters are positive integers, so we have:

$$I_{\frac{\beta}{t+\beta}}(\alpha N, n) \geq I_{\frac{\beta}{t+\beta}}(\alpha N + 1, n)$$

Thus, since all other parameters are positive integers, we use induction on N , so we have, for $t > 0$:

$$S_A(t) = I_{\frac{\beta}{t+\beta}}(\alpha N_1, n) \geq I_{\frac{\beta}{t+\beta}}(\alpha N_2, n) = S_B(t)$$

Therefore, the recruitment time of trial A with N_1 centers has stochastic dominance over that of trial B with N_2 centers. \square

Proposition 2. Assume that there are two clinical trials A and B , where trial A has the shape parameter value α_1 and trial B has the shape parameter α_2 ($\alpha_1 < \alpha_2$), and the other parameters are the same. The recruitment time of trial A has stochastic dominance over that of trial B .

Proof. From the proof of Proposition 1, we have:

$$I_{\frac{\beta}{t+\beta}}(\alpha N, n) \geq I_{\frac{\beta}{t+\beta}}(\alpha N + 1, n)$$

We use induction on α instead of N , and then we have

$$S_A(t) = I_{\frac{\beta}{t+\beta}}(\alpha_1 N, n) \geq I_{\frac{\beta}{t+\beta}}(\alpha_2 N, n) = S_B(t)$$

Therefore, the recruitment time of trial A with α_1 stochastic dominance over trial B with α_2 . \square

Proposition 3. Assume that there are two clinical trials A and B , where trial A needs to recruit n_1 patients and trial B need to recruit n_2 centers ($n_1 > n_2$), and the other parameters are the same. The recruitment time of trial A has stochastic dominance over trial B .

Proof. Like Proposition 1 and 2, we obtain the survivor function [9]:

$$S(t) = \frac{\beta^{\alpha N}}{\mathcal{B}(n, \alpha N)} \int_t^\infty \frac{x^{n-1}}{(x + \beta)^{n+\alpha N}} dx = I_{\frac{\beta}{t+\beta}}(\alpha N, n)$$

It is also well-known that [9]:

$$I_{\frac{\beta}{t+\beta}}(\alpha N, n) = I_{\frac{\beta}{t+\beta}}(\alpha N, n+1) - \frac{(\frac{\beta}{t+\beta})^{\alpha N} (\frac{t}{t+\beta})^n}{n \cdot \mathcal{B}(\alpha N, n)}$$

For any fixed $t > 0$, all other parameters are positive integers, so we have:

$$I_{\frac{\beta}{t+\beta}}(\alpha N, n+1) \geq I_{\frac{\beta}{t+\beta}}(\alpha N, n)$$

Thus, since all other parameters are positive integers, we use induction on N , so we have:

$$I_{\frac{\beta}{t+\beta}}(\alpha N, n_1) \geq I_{\frac{\beta}{t+\beta}}(\alpha N, n_2)$$

Therefore, the recruitment time of trial A recruiting n_1 patients has stochastic dominance over that of trial B recruiting n_2 patients. \square

Proposition 4. Assume that there are two clinical trials A and B , where trial A has

the rate parameter β_1 and trial B has the rate parameter β_2 ($\beta_1 < \beta_2$), and the other parameters are the same. The recruitment time of trial A has stochastic dominance over trial B .

Proof. We obtain the survivor function:

$$S(t) = I_{\frac{\beta}{t+\beta}}(\alpha N, n)$$

It is well-known that [9] for any fixed $t > 0$:

$$I_{\frac{\beta}{t+\beta}}(\alpha N, n) = B_{\frac{\beta}{t+\beta}}(\alpha N, n) / B(\alpha N, n)$$

$$B_{\frac{\beta}{t+\beta}}(\alpha N, n) = \int_0^{\frac{\beta}{t+\beta}} t^{\alpha N-1} (1-t)^{n-1} dt$$

Since α , β , N and n are positive integers, we have for any fixed $0 < t < 1$, $t^{\alpha N-1}(1-t)^{n-1} \geq 0$. Since $\beta_1 < \beta_2$, we also have $\frac{\beta_1}{t+\beta_1} < \frac{\beta_2}{t+\beta_2}$. Hence, we have the following:

$$B_{\frac{\beta_1}{t+\beta_1}}(\alpha N, n) = \int_0^{\frac{\beta_1}{t+\beta_1}} t^{\alpha N-1} (1-t)^{n-1} dt \leq \int_0^{\frac{\beta_2}{t+\beta_2}} t^{\alpha N-1} (1-t)^{n-1} dt = B_{\frac{\beta_2}{t+\beta_2}}(\alpha N, n)$$

$$S_A(t) = I_{\frac{\beta_1}{t+\beta_1}}(\alpha N, n) \leq I_{\frac{\beta_2}{t+\beta_2}}(\alpha N, n) = S_B(t)$$

Therefore, the recruitment time of trial A with β_2 has stochastic dominance over that of trial B with β_1 . \square

Now we also consider the special cases of fixed α/β , since it means fixing the mean rate of Poisson processes. With fixed n and N , the mean time of recruitment is almost equal because the mean is $\mu = E[T] = \beta n / (\alpha N - 1)$ [9]. Moreover, when having larger values of α and β with their ratio unchanged, the variance of the rates, α/β^2 , is lower, and as the probability density function is bell-shaped, we can observe stochastic

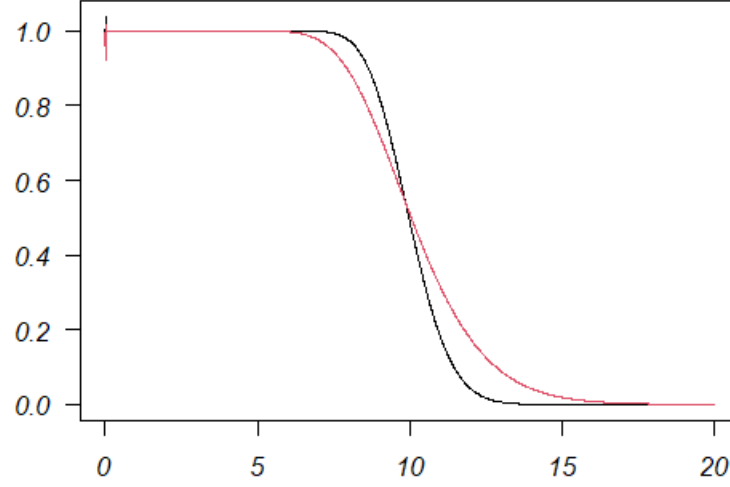


Figure 2.1: The survivor function of the recruitment time in a multi-site trial with $\alpha = 100, \beta = 50, n = 100, N = 5$ (black) and $\alpha = 8, \beta = 4, n = 100, N = 5$ (red).



Figure 2.2: On the left is the survivor function of the recruitment time in a multi-site trial with $\alpha = 100, \beta = 50, n = 130, N = 6$ (black) and $\alpha = 80, \beta = 40, n = 100, N = 5$ (red). On the right is the survivor function of the recruitment time in a multi-site trial with $\alpha = 100, \beta = 50, n = 130, N = 6$ (black) and $\alpha = 8, \beta = 4, n = 100, N = 5$ (red).

dominance of recruitment time with larger α and β over smaller α and β in t smaller than the mean and vice versa in t greater than the mean, like in Figure 2.1.

Moreover, although α/β is fixed, the stochastic order with different values of n and N can still be changed with enough differences between different values of α and

β . In Figure 2.2, although the black one has stochastic dominance over the red on the left, as α and β values become smaller, the dominance is overturned at a certain point as the variance decreases.

Chapter 3

Estimation Problems

In reality, we are never given the original gamma distribution from which the recruitment rates are generated, so it is hard to use the distribution we obtained in the last chapter. Instead, we can only obtain data with previous trials and recruitment time, so we want to predict the recruitment time with the given information using estimation methods.

3.1 Simulated Study

Specifically, we observe previous data of time T for recruiting n patients in N operating centers, and we want to obtain a reasonable estimate of rates to have more precise predictions. To achieve this goal, our estimation focuses on the parameters of the gamma distribution of the mean rates in independent Poisson processes, because a good prediction of it can make our prediction of mean recruitment rates in each center more precise, and they are essential to the recruitment time in each center.

First of all, we generate data from simulations. Since the rates are from a gamma distribution, we generate N samples from the gamma distribution with shape parameter α and rate parameter β , and we have now obtained the recruitment rates for the N centers. Now that each center has a Poisson process of patient arrivals, the most

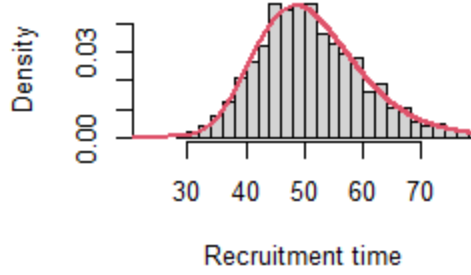


Figure 3.1: The PearsonVI distribution of T and the histogram of the simulated recruitment time with $\alpha = 10, \beta = 10, n = 200, N = 4$.

crucial step is to find the exact time that these centers recruit n patients. However, suppose we only approach the recruitment time by trying to determine the n -th patient in its center. In that case, it will be difficult to track which center provides the last arrival because people arrive randomly and at different rates across centers, so it is uncertain how many patients will be recruited in each center. Therefore, instead of determining the center of n -th patient, we simply let each center recruit n patients, guaranteeing that these centers recruit n patients in total. Using the fact that the time between counts of a Poisson process follows an exponential distribution[7], we can easily track the exact time of every recruited patient in each center. As a result, we can count the first n patients across centers and obtain the recruitment time of the last person, giving the exact recruitment time of n patients. In this way, we have created one simulated result of $T(\alpha, \beta, n, N)$.

To see the recruitment time follows the PearsonVI distribution[2], we plot the histogram of simulated results with the PearsonVI distribution with the cumulative density function of the corresponding distribution with the parameters that the results are generated from, which are displayed in Figure 3.1.

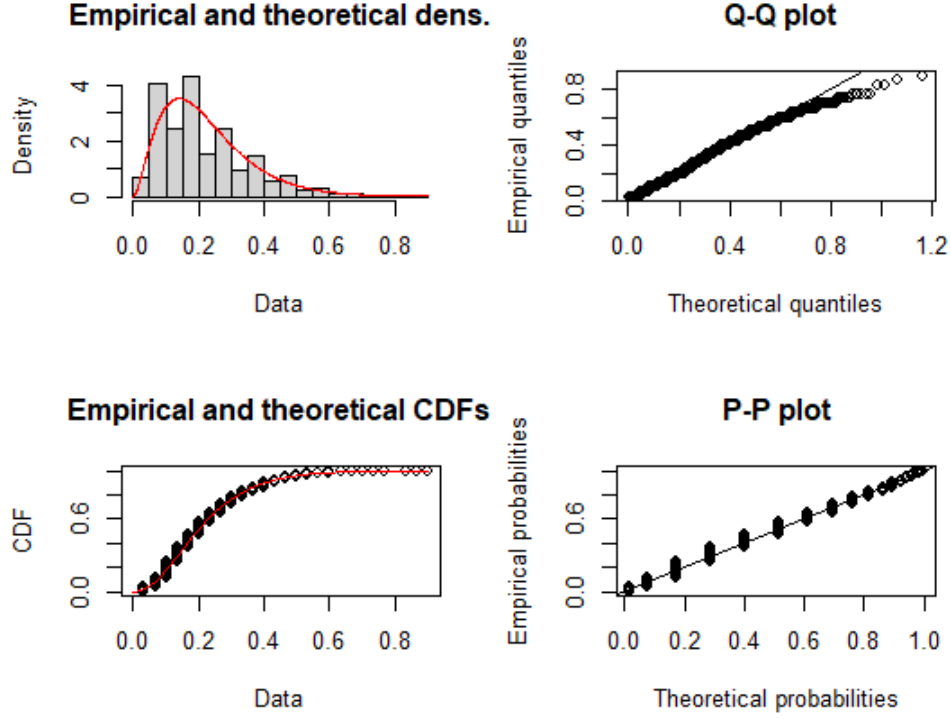


Figure 3.2: Simulations with $\alpha = 4$, $\beta = 20$, $N = 20$, $T = 30$. Fit into a gamma distribution and obtain $\hat{\alpha} = 3.883$ and $\hat{\beta} = 17.360$

3.2 Estimation methods

It is more conventional to obtain the number of people recruited in a given period instead of the specific time of recruiting n patients since the centers count their recruitment daily. Thus, we revised our code, which tracks the number of people recruited in a given interval, like 30 days.

Now, we use two approaches for the estimation. In the first one, we treat all the data generated as the mean rates. To be more specific, for each recruitment time of simulation, we divide it by the length of time interval (like 30 days), so now we have the mean recruitment rate. We use them as the rates from the original gamma distribution to estimate the original distribution of rates, so we fit them directly into a gamma distribution to find the estimated parameters $\hat{\alpha}$ and $\hat{\beta}$ (See Figure 3.2).

Alternatively, according to the theoretical results [12], we can estimate the problem

using maximum likelihood estimation to find the best parameters.

Definition 6 (Maximum likelihood estimation). Maximum likelihood estimation (MLE) estimates the parameters of an assumed probability distribution, given some observed data.[10] This is achieved by maximizing a likelihood function so that the observed data is most probable under the assumed statistical model. The likelihood function is given by:

$$\mathcal{L}(\theta|x) = \prod_{i=1}^n f(x_i|\theta),$$

where $\theta \in \Theta$, the set of parameters, $\mathbf{x} = \{x_1, x_2 \dots x_n\}$ is the set of observed data values, and $f(x_i|\theta)$ is the probability density function of observed data value for x_i in \mathbf{x} , given its parameter value θ . The maximum likelihood estimator (MLE) for the set of parameters Θ is obtained by maximizing the likelihood function:

$$\hat{\Theta}_{\text{MLE}} = \arg \max_{\Theta} \mathcal{L}(\Theta|\mathbf{x})$$

In our model, according to [8], we may derive the maximum likelihood function for our recruitment time. Consider λ_c , the mean rate of recruitment for center c drawn from a $\text{Gamma}(\alpha, \beta)$ function. In a given time interval t , like 30 days, we assume that there is a recruitment of n_c^s people for each center c on time s , so the total recruitment and center c is $n_c = \sum_{s=1}^t n_c^s$ and the total recruitment $n = \sum_{c=1}^N n_c$. Hence, we have for center c , from time 1 to t , the probability density function of n_c^s is a Poisson variable for time s in $1 : t$, so we write the probability density function of data values of the number of recruitment n_c^s as

$$\frac{\lambda_c^{n_c^s}}{n_c^s!} \exp(-\lambda_c)$$

However, λ_c is not the coefficient we want to estimate, and it follows the $\text{gamma}(\alpha, \beta)$ distribution, so we use the continuous mixture model in [7] to integrate.

Definition 7 (Continuous mixtures). Assume that the random mixing parameter has a continuous distribution, then:

$$f_T(t) = \int_{-\infty}^{\infty} f_{T|\Theta=\theta}(t|\Theta = \theta) f_{\Theta}(\theta) d\theta,$$

where:

- θ is the random parameter.
- $f_{\Theta}(\theta)$ is the distribution of the random parameter.
- $f_{T|\Theta=\theta}(t|\Theta = \theta)$ is the conditional distribution of time given the value of θ .

In our case, we have the distribution of the random parameter $\lambda_c > 0$:

$$f_{\Lambda}(\lambda_c) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda_c^{\alpha-1} \exp(-\beta \lambda_c)$$

We also have the conditional distribution of time given the value of λ_c :

$$f_{T|\Lambda=\lambda_c}(t|\Lambda = \lambda_c) = \prod_{s=1}^t \frac{\lambda_c^{n_c^s}}{n_c^s!} \exp(-\lambda_c)$$

The likelihood function for center c is given by:

$$\begin{aligned} L(\alpha, \beta; n_c^{1:t}) &= \int_0^{\infty} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda_c^{\alpha-1} \exp(-\beta \lambda_c) \prod_{s=1}^t \frac{\lambda_c^{n_c^s}}{n_c^s!} \exp(-\lambda_c) d\lambda_c \\ &\propto \frac{\beta^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} \lambda_c^{\alpha+n_c-1} \exp[-\lambda_c(\beta+t)] d\lambda_c = \frac{\Gamma(\alpha+n_c)}{\Gamma(\alpha)} \frac{\beta^{\alpha}}{(\beta+t)^{\alpha+n_c}} \end{aligned}$$

Hence, according to Definition 6, up to a constant, take the logarithm of the product of the likelihood functions among all centers, we obtain the log-likelihood function of

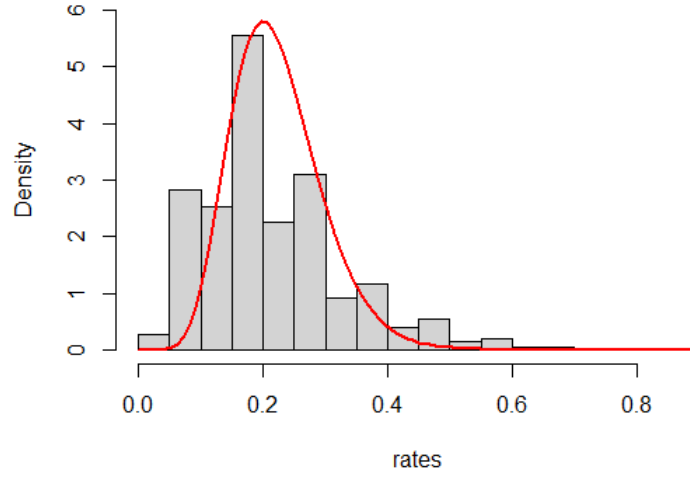


Figure 3.3: Simulations with $\alpha = 4$, $\beta = 20$, $N = 20$, $T = 30$ and using optimization to obtain fitted parameters with $\hat{\alpha} = 9.647$ and $\hat{\beta} = 43.122$

the collective data across all centers:

$$\ell(\alpha, \beta) = N\alpha \log \beta - (N\alpha + n) \log (\beta + t) - N \log \Gamma(\alpha) + \sum_{c=1}^N \log \Gamma(\alpha + n_c).$$

Thus, by maximizing this likelihood function, we can find the maximum likelihood estimator for the parameters α and β . To realize this goal, we implement this function as the objective function in an optimization solver **Optimr** to obtain the maximum likelihood estimator for parameters α and β from the same data that we fitted directly into a gamma distribution in Figure 3.2. The fitted gamma distribution of rates is in Figure 3.3.

3.3 Optimization Problem

In the last section, we discussed the log-likelihood function to be maximized. We now convert the maximization to an optimization problem of minimization. Since minimizing the negation is the same as maximizing the original log-likelihood function,

by using the negation of this function as the objective function, we can use the solver to find $\hat{\alpha}$ and $\hat{\beta}$ that minimize the objective function and hence become the maximum likelihood estimator.

Chapter 4

Numerical Studies

In this chapter, we will implement the optimization problem in the last chapter and compare the performance of different algorithms on the estimation of recruitment time.

4.1 Performance of Optimization Algorithms

To find the optimal solution for the optimization problem, we use different optimization algorithms in **Optimr** and compare their performance in estimating the recruitment time $T(n, N, \alpha, \beta)$. Since we can obtain the number of patients n and the number of centers N , α and β are the only parameters we are missing.

Firstly, we generate simulated data and compare the estimated parameters from all algorithms by plotting their recruitment time distribution. The fitted parameters are usually very different from the original ones during experiments. Figure 4.1 shows that some algorithms, including nonlinear minimization (NLM), nonlinear conjugate gradient minimization (RCGMIN), variable metric nonlinear function minimization (RVMMIN), nonlinear minimization using boundary constraints (NLMINB), find a local minimum and stop, generally stopping within a small range of values from the initial parameters. For instance, NLM stops not only at $\hat{\alpha} = 2.42$ and $\hat{\beta} = 2.187$ with

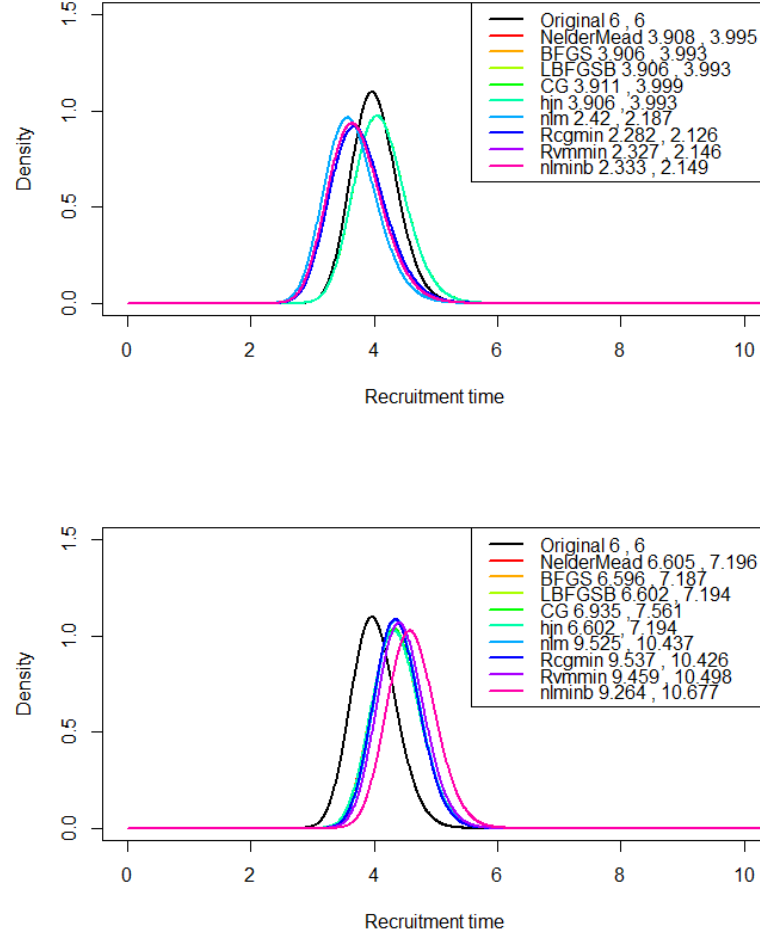


Figure 4.1: Simulations with $\alpha = 6$, $\beta = 6$, $N = 50$, $n = 200$, starting with initial parameters $\alpha = \beta = 2$ on the top and $\alpha = \beta = 10$ on the bottom. The numbers behind each name of optimization are $\hat{\alpha}$ and $\hat{\beta}$.

initial $\alpha = \beta = 2$ but also at $\hat{\alpha} = 9.525$ and $\beta = 10.437$ with initial $\alpha = \beta = 10$, which are far from the actual values.

On the other hand, we look at the algorithms that have comparably better performances: the Nelder–Mead method, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, a variant of BFGS (L-BFGS-B), conjugate gradients (CG), and Hooke and Jeeves Pattern Search (HJN). From Figure 4.1, We compare them with the previous ones, finding these methods earn more favorable results closer to the true values of parameters. To further examine the underlying factors of the differences between those

methods with better estimations and those with unfavorable ones, we take the nonlinear minimization (NLM) from the former and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) from the latter, and we compare them by looking closely into their algorithms. We introduce the following definitions to help us understand how they work:

Definition 8 (Hessian matrix). The Hessian matrix of a scalar-valued function f of n variables $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is a square matrix of second partial derivatives of f . It is denoted by \mathbf{H} or $\nabla^2 f$, and its elements are given by:

$$\mathbf{H}(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Definition 9 (Gradient). The gradient of a scalar-valued function f of n variables $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is a vector of its first partial derivatives. It is denoted by ∇f or $\frac{\partial f}{\partial \mathbf{x}}$, and its components are given by:

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)$$

Now we compute the gradient and Hessian matrix of the original function in our case:

$$\begin{aligned} \nabla \ell(\alpha, \beta) &= \left(\frac{\partial \ell}{\partial \alpha}, \frac{\partial \ell}{\partial \beta} \right) \\ &= \left(N \log(\beta/(\beta + t)) - N\psi(\alpha) + \sum_{c=1}^N \psi(\alpha + n_c), N\alpha/\beta - (N\alpha + n)/(\beta + t) \right) \end{aligned}$$

$$\mathbf{H}(\alpha, \beta) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \alpha^2} & \frac{\partial^2 \ell}{\partial \alpha \partial \beta} \\ \frac{\partial^2 \ell}{\partial \beta \partial \alpha} & \frac{\partial^2 \ell}{\partial \beta^2} \end{bmatrix}$$

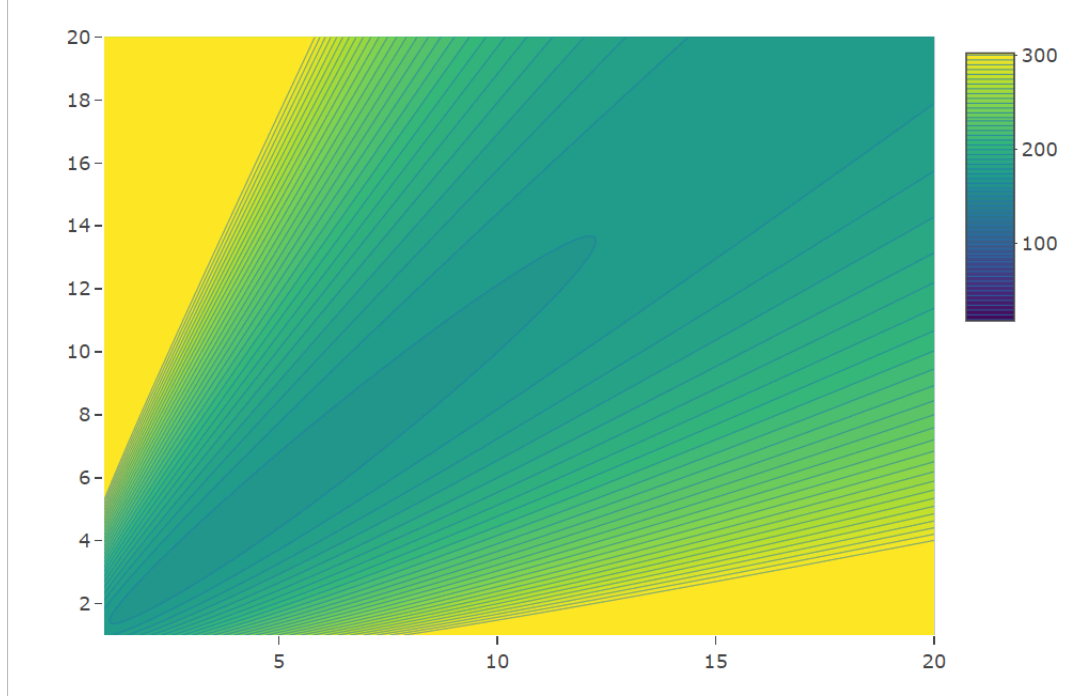


Figure 4.2: Contour map of the objective function with α and β from 1 to 10 (simulations with $\alpha = 6$, $\beta = 6$, $N = 50$, $n = 200$).

$$= \begin{bmatrix} -\sum_{c=1}^N 1/n_c & Nt/\beta(\beta+t) \\ Nt/\beta(\beta+t) & -N\alpha/\beta^2 + (N\alpha+n)/(\beta+t)^2 \end{bmatrix}$$

Where ψ is the digamma function where $\psi(x) = \Gamma'(x)/\Gamma(x)$ with the gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$.

A potential explanation for the unsatisfactory performance of the nonlinear minimization(NLM) is extremely sensitive to the shape of the objective function, so it converges very fast. But in this problem, the values of the function are very close to the minimum in a large area of estimated parameters from the contour map in Figure 4.2. In other words, the nonlinear minimization finds a value quickly because it is already very close to the minimal value of the function, so it stops within a close range of initial parameters.

4.2 Robustness of Algorithms

The previous section suggests the favorable methods are the Nelder–Mead method, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, a variant of BFGS (L-BFGS-B), conjugate gradients (CG), and Hooke and Jeeves Pattern Search (HJN). Next, we will study their robustness compared with those that are not favorable, given changes in some parameters to the optimization.

Starting points. Firstly, we changed the starting parameters to observe their behaviors (see Figure 4.1). We changed the starting point from $\alpha = \beta = 2$ to $\alpha = \beta = 10$, and the $\hat{\alpha}$ and $\hat{\beta}$. As mentioned before, the unfavorable methods earned a result within a close range of initial parameters. On the other hand, the favorable methods earned a less divergent result from the true values. For instance, BFGS earned $\hat{\alpha} = 3.906$ and $\hat{\beta} = 3.993$ starting with $\alpha = \beta = 2$ and $\hat{\alpha} = 6.596$ and $\hat{\beta} = 7.187$ starting with $\alpha = \beta = 10$, and they are much closer to the true values of $\alpha = \beta = 6$ than the results in the unfavorable methods. Therefore, the unfavorable methods are more sensitive to starting points than others, so the starting point has less impact on the favorable methods so that they can give a better estimation.

Number of sites. Our numerical experiments suggest that the optimization methods depend on the number of recruitment sites. Keeping the other parameters fixed, we changed the centers from 50 to 20 to 5 and ran the simulation several times for each case to see the changes in results. The distribution of $\hat{\alpha}$ and $\hat{\beta}$ are plotted in Figure 4.3. We observed that the distributions of $\hat{\alpha}$ and $\hat{\beta}$ become more and more flat when the number of centers decreases, with greater variance.

A potential explanation for this phenomenon can be derived from our estimation methods of the Poisson-gamma model and the simulation process. In each center, we generated a Poisson process. We used the number of recruited patients from one center to estimate the mean recruitment rate in that center. Hence, the more centers we have, the better we can estimate the values of α and β because we now have observed

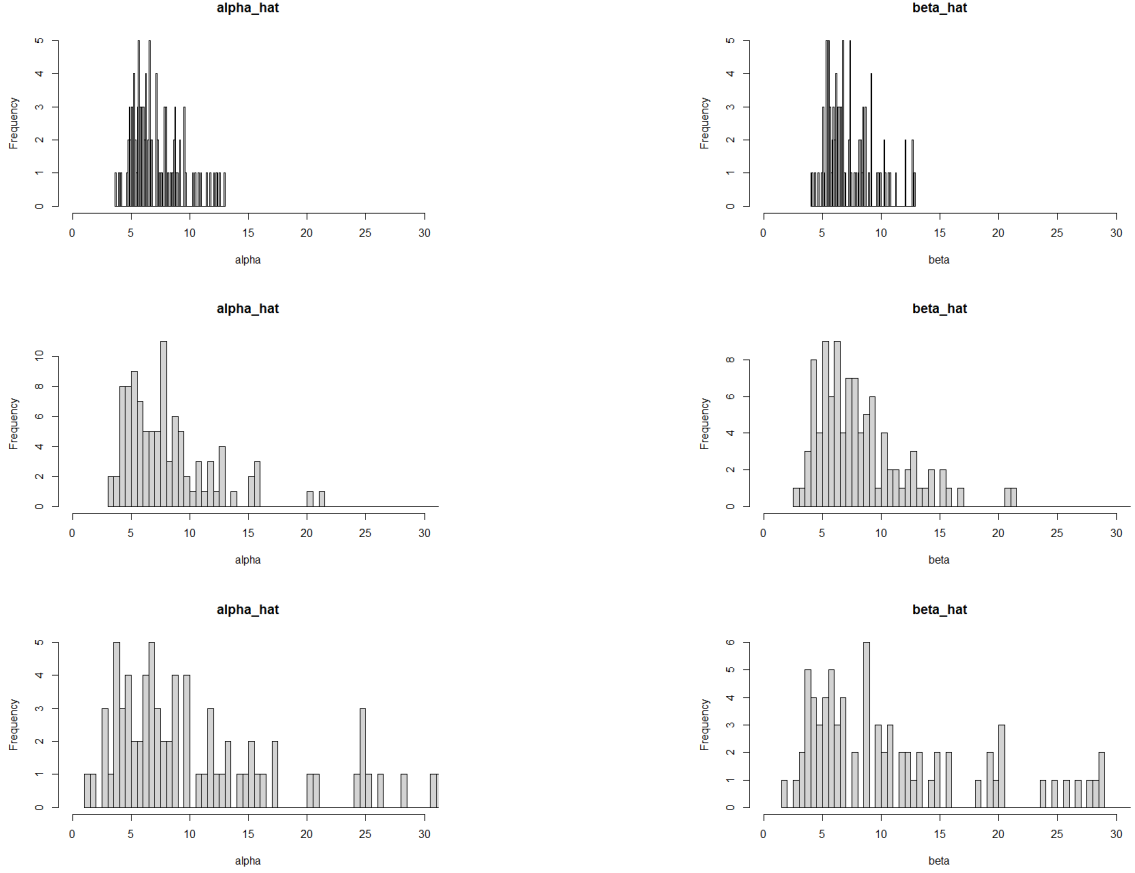


Figure 4.3: Distributions of $\hat{\alpha}$ and $\hat{\beta}$ with simulations from $\alpha = 6$, $\beta = 6$, $n = 200$, $N = 50$, 20, 5 on the first, second and third row respectively.

more values from the gamma distribution of mean rates.

For a numerical and theoretical explanation for this phenomenon, we take the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm as an example to illustrate it. The accurateness of his method relies on the convexity of the problem, which is satisfied when the Hessian matrix is positive definite [4].

Recall in our case, we are trying to minimize the negation of the function, so the Hessian matrix of the optimization problem is the negation of the Hessian matrix of the log-likelihood function, as follows:

$$\mathbf{H} = \begin{bmatrix} \sum_{c=1}^N 1/n_c & -Nt/\beta(\beta + t) \\ -Nt/\beta(\beta + t) & N\alpha/\beta^2 - (N\alpha + n)/(\beta + t)^2 \end{bmatrix}$$

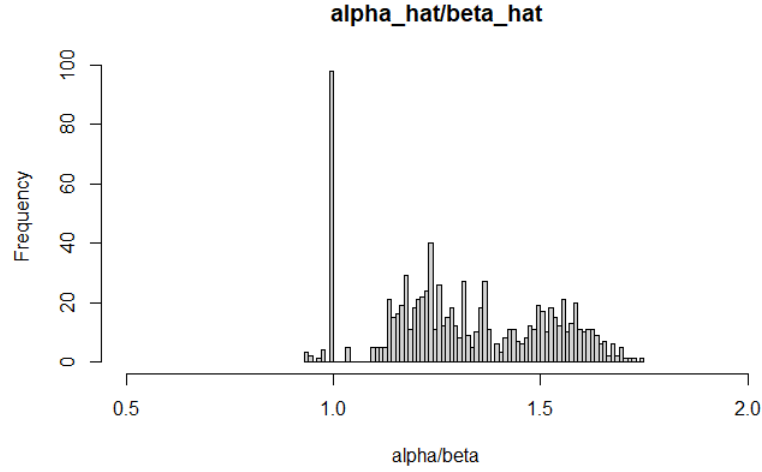


Figure 4.4: Histogram of $\hat{\alpha}/\hat{\beta}$ with simulations from $\alpha = 6$, $\beta = 6$, $N = 50$, $n = 200$).

From [6], the positive definiteness is satisfied when the trace and determinant of this matrix are positive. They both increase quadratically when we increase N , resulting in a greater possibility of making the matrix positive definite, guaranteeing the convexity.

Notably, even though we increased the number of sites to 50, from the distributions of $\hat{\alpha}$ and $\hat{\beta}$ in Figure 4.3, there are still variations in the estimations. However, the estimate of $\frac{\alpha}{\beta}$, the mean of the recruitment rates, is quite stable. In Figure 4.4, $\hat{\alpha}/\hat{\beta}$ stays in a close range from 1 to 1.5, with the original $\alpha/\beta = 6/6 = 1$.

Chapter 5

Conclusion and Future Research

In this thesis, we studied the properties of recruitment time distribution of the Poisson-gamma model and implemented the model to our estimation. From numerical results, the estimation of this model from simulated data is not stable using the optimization packages in R, and a potential explanation for this problem is the small number of centers. However, repetitions of simulations also show that the favorable optimization methods are still unstable, so there may be some endogenous problems to be discovered in this model. For example, we observe the fact that these estimated parameters of α and β give a mean that is close to the original gamma distribution, and this leads to a question of whether using the gamma function is necessary since more centers will lead to the convergence of their mean rates to the mean of the gamma distribution by the Central Limit Theorem. In contrast, a small number of centers may not require the gamma distribution since a fixed mean may be already enough for a good estimation, and fitting a small data of mean rates into a gamma distribution may even cause more errors instead of using individual values directly.

Meanwhile, the Poisson-gamma model is helpful in many other applications, such as ecology. In our case, the α and β values are usually small but may be significant in these models. Therefore, when optimizing the maximum likelihood function, we may

counter the problem that the gamma function will expand exceptionally quickly when these parameters are large, causing trouble in the optimization process. Therefore, there may be some more innovative ways to solve this problem using other methods.

Notably, there are more complicated models based on ours, like the ones with time-varying rates (non-homogeneous Poisson process, which is also well-developed in reliability) or those with different starting times and census times, which may give more perspectives on the estimation. By having reasonable estimations of recruitment rates across centers, people can predict enrollment in the future and help design experiments more efficiently.

Appendix A

Appendix

Bibliography

- [1] Angel G Angelov, Magnus Ekström, Bengt Kriström, and Mats E Nilsson. Four-decision tests for stochastic dominance, with an application to environmental psychophysics. *Journal of Mathematical Psychology*, 93:102281, 2019.
- [2] Vladimir V Anisimov and Valerii V Fedorov. Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in medicine*, 26(27):4958–4975, 2007.
- [3] Kirsty Barnes. Pharma giants risk reputation through clinical trial cost-cutting. *Accessed on October, 2, 2006*.
- [4] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [5] Rickey E Carter, Susan C Sonne, and Kathleen T Brady. Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC medical research methodology*, 5:1–5, 2005.
- [6] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.
- [7] Lawrence M Leemis. *Reliability: probabilistic models and statistical methods*. Prentice-Hall, Inc., 1995.

- [8] Rachael Mountain and Chris Sherlock. Recruitment prediction for multicenter clinical trials based on a hierarchical poisson–gamma model: Asymptotic analysis and improved intervals. *Biometrics*, 78(2):636–648, 2022.
- [9] F.WJ Olver, D.W Lozier, R.F Boisvert, and C.W Clark. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.
- [10] James P Quirk and Rubin Saposnik. Admissibility and measurable utility functions. *The Review of Economic Studies*, 29(2):140–146, 1962.
- [11] Stephen Senn. Some controversies in planning and analysing multi-centre trials. *Statistics in medicine*, 17(15-16):1753–1765, 1998.
- [12] Szymon Urbas, Chris Sherlock, and Paul Metcalfe. Interim recruitment prediction for multi-center clinical trials. *Biostatistics*, 23(2):485–506, 2022.