新冠疫苗有效性试验设计及统计分析

论文题目:新冠疫苗有效性试验设计及统计分析

参赛学校:厦门大学

参赛成员(作者):王帅、李子恒、徐晓宇

指导老师:林建华

目录

表格和插图清单

摘要

新冠肺炎疫情给全球人类健康带来了前所未有的威胁。截至目前，全球新冠肺炎确诊人数已超过 1.4 亿，300 多万人死亡。在各国疫情此起彼伏、疫情形势充满变数的背景下，疫苗研发被寄予厚望。现如今各国都在积极研发新冠疫苗，其安全性及有效性成为了人们最关心的问题。

本文立足于这一现实背景，设计一些具有现实意义的试验，通过统计模拟临床试验的缺失数据，基于“ 0 ”点有概率的混合分布模型，分析疫苗失效时间在不同分布下的汇总数据，进而总结出疫苗的失效时间用哪种分布函数模拟比较合适。从防疫角度上来看，当知道了疫苗失效时间的分布情况，可以很快的评估某个时间段内还有多少人体内的抗体是有效的。本文的研究目的是探索这种统计分析的可行性，这对疫苗有效性的研究有着重要的借鉴意义。

论文主要尝试通过模拟服从指数分布和 Pareto 分布的右截断缺失数据及左缺失数据，采用最大似然估计法列出相应的最大似然函数，求出对数似然函数及其一二阶导数，进而求解出最大似然估计及其方差，最后通过似然比检验来验证假设的可行性及合理性。

结果显示，带缺失数据时的最大似然估计值与真实值的拟合程度较好，经过对比，右截断缺失数据的分布情况与 Pareto 分布较为近似，左缺失数据的分布情况也与 Pareto 分布相近。所以，从总体上看，试验设计合理的情况下，疫苗的失效时间可以采用 Pareto 分布来模拟。

关键词：缺失数据；疫苗有效性；抗体浓度；最大似然估计；似然比检验

Abstract

The COVID-19 outbreak poses an unprecedented threat to global human health. To date, more than 140 million people worldwide have been diagnosed with COVID-19, resulting in more than 3 million deaths. High hopes have been placed on the research and development of vaccines against the backdrop of ever-changing epidemics in various countries. The safety and efficacy of the new coronavirus vaccine have become the most concerned issue in the current development of the new coronavirus vaccine.

Based on the reality of the background, design some realistic test, through simulation the lack of clinical trial data statistics, based on the "0" mixed distribution model of probability, the analysis of the vaccine failure time under different distribution of summary data, and then sums up the vaccine failure simulation more appropriate time which one to use distribution function. From a prevention point of view, knowing the distribution of vaccine failure time allows you to quickly assess how many antibodies are still available in a given period of time. The purpose of this study is to explore the feasibility of this statistical analysis, which has important reference significance for the study of vaccine effectiveness.

Thesis mainly attempt by simulating obeys exponential distribution and Pareto distribution right truncation missing data and left missing data, using the maximum likelihood estimation method lists the corresponding maximum likelihood function, and the logarithmic likelihood function and a second derivative, thus solving the maximum

likelihood estimation and variance, finally through the likelihood ratio test to verify the feasibility and rationality of assumption.

The results show that the maximum likelihood estimated value with missing data is in good fit with the true value. After comparison, the distribution of missing data with right truncation is similar to that of Pareto distribution, and the distribution of missing data with left truncation is similar to that of Pareto distribution. Therefore, in general, if the test design is reasonable, the failure time of vaccine can be simulated by using the Pareto distribution.

Keyword: Missing data; Vaccine effectiveness; Antibody concentration; Maximum likelihood estimation; likelihood ratio test

一、引言

1.1 研究背景

2020 年对所有人来说，无疑是人生中最特殊的一年，一场突如其来的疫情，措手不及的改变了我们对这个世界的认知。在中国共产党的伟大领导下，中国的疫情取得了重大成果，但时至今日新冠疫情仍在大部分国家肆虐。中国工程院院士钟南山曾经说过：“疫苗是解决疫情最根本的东西，最终形成群体免疫是靠疫苗。”所以新冠疫苗的开发工作是现在最重要的事情，是国之重器。而对于普通民众来说，最关心的问题莫过于疫苗的有效性问题。这也是本次论文研究的主要方向。

评估一款疫苗通常有三个步骤——安全性，免疫原性和保护效力，分别对应了临床研究中的 I，II 和 III 期阶段。而在 I 期安全性临床完成后，II/III 期的一个重点内容就是对疫苗的有效性进行研究。虽然 II 期临床中测得的各类免疫反应数据不能直接翻译成最终保护效率，但也无法否认两者之间存在关联。

4 月 15 日，Moderna 疫苗日活动（Moderna 2nd Annual Vaccines Day）上，新南威尔士大学柯比研究所（Kirby Institute）传染病分析项目主管 Miles P Davenport 教授分享了疫苗免疫原性数据和疫苗保护效率数据之间存在怎样的关联和这种关系带来的启示。

虽然不同疫苗的临床设计不同，检测手段各异，导致疫苗的中和抗体数据无法直接比较。但接种疫苗后，受试者产生的中和抗体滴度相比同一研究中测得的康复患者血清滴度的比值是一个可行的标准化方案。例如，Moderna 新冠 mRNA 疫苗 I 期临床研究测得受试者中和抗体滴度为 654.3，而康复患者血清抗体滴度为 158.3，将血清抗体滴度视为 1，那么该疫苗的中和抗体比值为就为 3.8，意味着疫苗诱导的抗体水平是自然感染康复患者血液内抗体水平的 3.8 倍。

根据现有的研究数据，不难发现，随着时间推移，相同时间段内，产生更高中和抗体滴度的疫苗依然对普通与重症具有预防作用，而起始中和抗体滴度较低的疫苗可能在较短时间内就失去对普通甚至重症的防护作用。例如，中和抗体比值 4 的疫苗，到第 3 年都可以保持 50%以上的保护效率，而中和抗体比值 0.5 的疫苗可能在半年内就无法达到合格的保护效力，这也意味着可能需要增强免疫或更频繁的接种。

这些实验有着一定的参考意义，但很显然都有一个明显的弊端，那就是必须要在有疫情发生的国家进行临床试验。因为在疫情发生的国家才能拿到大量感染新冠病毒患者的数据，而现如今中国抗疫效果显著，国内已几乎没有确诊病例（除输入病例），所以有关对照组的实验数据很难拿到。此外，抗体在体内每时每刻都在发生着变化，想要拿到每个人抗体浓度的实时变化数据是不可能的。对此我们可以模拟缺失数据，并在接下来的分析中探究其是否可行。

1.2 研究方法及意义

本文以接种疫苗后抗体浓度产生的峰值时间作为出发点，通过查阅相关文献及资料，假设一个合理的峰值，并以此建立一个混合分布模型。以峰值出现的时间为“0”点，用最大似然估计法求出“0”点的估计，从而将模型简化为连续性分布模型。在考虑连续性分布时，假设总体服从指数分布和 Pareto 分布，并用 R 软件模拟出相应区间的汇总数据，即疫苗失效时间的数据。此时需要模拟的是较为复杂的环境背景下的数据，比如右截断的缺失数据（“带 right-censored”的缺失数据）及左缺失数据（带“left-truncated”的缺失数据）。进而用最大似然估计法求出参数的估计值，与模拟的真实值对比之后，分析参数的拟合程度，再用 Fisher 信息量计算最大似然估计值的方差，最后用似然比检验验证指数分布和 Pareto 分布的合理性。

本文采用的实验数据为缺失数据，不包含真实数据，所以可以认为是一种实验设计，目的在于给其他人一个可以快速入手的框架，希望能够为以后研究疫苗有效性提供相应的参考。

论文的结构主要如下：第一部分为论文的研究背景、研究方法及意义；第二部分为试验设计；第三部分为模型的建立与求解过程；第四部分为假设检验；第五部分为总结与建议。

二、试验设计

现在全国范围内正在接种第二针疫苗，中国疫控中心免疫规划首席专家王华庆表示，在接种疫苗第二剂次 14 天后，会产生较好的免疫作用。在新冠疫苗的各期临床试验中也已经证实，接种新冠疫苗以后，机体多能产生一定滴度的保护性抗体，抗体水平约两周后将达到高峰，但随着接种时间的延长，机体内的保护性抗体滴度也会逐渐的降低甚至消失，其时间多在半年到一年左右时间。

本文的试验设计初衷是要尽可能简便一些，并且要贴合实际情况，所以我们可以假设接种第二针疫苗 14 天后抗体浓度达到最大，比如 300μg/ml，此后开始呈现下降趋势，直到最后失效。对此，我们以缺失数据来模拟疫苗的失效时间。并在此背景下，模拟一些不同的试验设计。

2.1 右截断缺失数据背景下的试验设计

我们希望样本量足够大，所以假设现在有 10100 名实验体，其中有 100 人在接种第二针后 14 天内抗体就失效了，或者说这一部分人没产生抗体，为了模型的简便，我们暂时不考虑这部分人。我们将第十四天作为起始时间，即“ 0 ”点，将其余人在这一点处做第一次的抽血检查，观察他们血液中抗体浓度的变化。由于个人体质以及外界环境因素等，这时候会有一小部分人抗体失效，我们记录下他们的实验数据，具体的实验数据可以在模型建立过程中由缺失数据假设。同时为了能更符合现实背景，我们希望所有人的抽血时间能间隔短一些，这样能保证可以在现实背景下实现，并且也能拿到这样的数据。所以在“ 0 ”点后的第 30 天，再次抽血检查，并记录此时的实验数据。再之后的 30 天，再次记录实验数据。到第 90 天时做最后一次的抽血检查，并记录实验数据。这时我们就得到了 $[0,30]$，$[30,60]$，$[60,90]$，90^{+}这四个区间的汇总数据，其中 90^{+}的数据就是右

截断的缺失数据（即带“ right-censored ”的缺失数据）。通过 R 软件模拟了以上四个区间的缺失数据，其总体分布服从单参数的指数分布以及 Pareto 分布。在此基础上，我们来分析疫苗失效时间的分布情况。

2.2 左缺失数据背景下的试验设计

左缺失数据与右截断的缺失数据情形差别较大。所以我们可以假设如下情景：目前我在做的实验在社会上的认可度较高，所以在实验进行到某一阶段时，有一部分人加入到了当前的实验中。这些人都是打完第二针疫苗后来进行抽血检查的，来检测自己体内的疫苗是否还有效。但此时的情况比较复杂，因为这些人打完第二针疫苗后来检测的时间不一样，比如过了几十天或几百天后才来检测。此时为了模型的简便，我们可以假设一个条件，必须要过了 114 天的人才可以来检测，并且此时还要确保他们的第一次测试是有效的，即体内的抗体还在起作用。无效的话就暂不列为实验对象，因为干扰因素较多，不清楚他们是在一开始体内的疫苗就没有起作用，还是在打完疫苗后某个时间点突然失效，所以为了规避这个问题，我们只选取过了 114 天还有效的实验体的数据，这时可以拿到很多左缺失的数据，即带“ left-truncated ”的缺失数据。而此时我们还是将 14 天作为初始时间，即一开始的“ 0 ”点，实验体仍为 10000 个。此时区间变为：[100, 130]，[130,160]，[160,190]，190^{+}。其中 190^{+}仍为右截断的缺失数据。再通过 R 软件模拟这四个区间的缺失数据。其总体分布依旧服从单参数的指数分布以及 Pareto 分布，在此基础上，我们再来分析此时疫苗失效时间的分布情况。

三、模型的建立与求解过程

针对第二节的两个试验，我们初步尝试用连续型分布模型进行求解。但由于假设的背景中存在一些特殊的数据，所以我们先考虑“0”点有概率的混合分布模型，将“0”点的估计求出之后，其余部分简化为连续型分布，再进行求解。首先要求出连续型分布的分布函数。 在“0”点的失效时间 Y 服从两点分布，$P(Y=0)=P_0$, $P(Y=1)=1-P_0$。即：

$$Y \sim \begin{pmatrix} 0 & 1 \\ P_0 & 1-P_0 \end{pmatrix}$$

大于“0”点的失效时间 X 为连续型分布，初步设为指数分布。即:X~exp(λ)

所以混合分布变量 Z=XY，其中 $P(Z=0)=P(Y=0)=P_0$，$f(Z=z)=P(Y=1)\cdot f(X=z)=(1-P_0)\lambda e^{-\lambda z},\ z>0$。$\lambda e^{-\lambda z}$为指数分布的概率密度函数。

综上，我们可以得到混合分布的分布函数为:

$$F(z)=\begin{cases} 0, & z<0 \\ P_0, & z=0 \\ (1-P_0)\lambda e^{-\lambda z}, & z>0 \end{cases}$$

同理，若 X 服从 Pareto 分布，则此时的混合分布的分布函数为：

$$F(z)=\begin{cases} 0, & z<0 \\ P_0, & z=0 \\ (1-P_0)\cdot \alpha\theta^{\alpha}/(z+\theta)^{\alpha+1}, & z>0 \end{cases}$$

$f(x)=\alpha\theta^{\alpha}/(x+\theta)^{\alpha+1}$为 Pareto 分布的概率密度函数。

其中“0”点的估计我们可以采用最大似然估计的方法。其计算方法为：

$$P(0)=\frac{\text{失效个体数}}{\text{总人数}}=\frac{100}{10100}=0.0099$$

现在余下部分为连续性分布，我们用连续性分布来解决两个试验的缺失数据问题。

3.1 试验一的建模与求解

3.1.1 指数分布的建模与求解

先解决第一个试验的缺失数据问题，用 R 软件产生指数分布随机数，其中参数λ=0.002，期望值 E(x) = 1/0.002 = 500 天，即平均失效天数为 500 天，方差 Var(x) = 1/(500×500) = 0.000004。我们得到了如下区间的汇总数据：

表 1　指数分布右截断的汇总数据

天数	（0，30）	（30，60）	（60，90）	90^{+}
抗体失效的实验体数	622	558	504	8316

接下来的求解我们采用最大似然估计的方法，其一般求解步骤如下所示：

1. 写出似然函数。似然函数的公式主要分以下几种情形：

①.对于连续性分布，其似然函数公式如下：

$$L(\theta) = f(x_1,\theta)\cdot f(x_2,\theta)\cdot f(x_3,\theta)\cdot \ldots \cdot f(x_n,\theta)$$

其中，$f(x_i,\theta)$为分布函数的密度函数。

②.对于离散型分布，其似然函数公式如下：

$$L(\theta) = L(\theta; x_1, x_2, \ldots, x_n) = p(X_1 = x_1|\theta)\cdot p(X_2 = x_2|\theta)\cdot p(X_3 = x_3|\theta)\cdot \ldots \cdot p(X_n = x_n|\theta)$$

其中，$p(X_i = x_i|\theta)$为离散性分布的概率密度。

③．对于混合分布模型，其似然函数公式如下：

$$L(\theta) = f(x_1,\theta)\cdot f(x_2,\theta)\cdot f(x_3,\theta)\cdot \ldots \cdot f(x_k,\theta)\cdot p(X_j = x_j)\cdot \ldots \cdot p(X_n = x_n)$$

其中，x_1，x_2，...，x_k要满足$f(x_k)$>0，i=1,....,k,$x_j, \ldots, x_n$要满足 $P(X_z = x_z)$>0，z =j,..n。

④．若分布函数含有左缺失及右截断的缺失数据，即对于所有的 i，有X_i>d，并且在 u 点右截断，那么此时的似然函数可以写成如下形式：

$$\mathrm{L}(\theta)=\frac{f(x_1,\theta)\cdot f(x_2,\theta)\cdot f(x_3,\theta)\cdot\ldots\cdot f(x_i,\theta)\cdot \mathrm{p}(X_j>\mathrm{u})}{[1-F(d)]^2}$$

其中，x_i≤u，$\mathrm{X_i}$ 在x_i点处右截断，此时考虑的即为随机变量[x|x>d]的分布。

2．对似然函数取对数，得到对数似然函数：

$$\ln[\mathrm{L}(\theta)]=\ln[f(x_1,\theta)\cdot f(x_2,\theta)\cdot f(x_3,\theta)\cdot\ldots\cdot f(x_n,\theta)]$$

3. 将对数似然函数整理后求其导函数，得到如下名为似然方程的方程：

$$\mathrm{l}(\theta)=\{\ln[f(x_1,\theta)\cdot f(x_2,\theta)\cdot f(x_3,\theta)\cdot\ldots\cdot f(x_n,\theta)]\}'$$

4．令似然方程等于 0，可求出关于 θ 的最大似然估计：

$$\mathrm{l}(\theta)=\{\ln[f(x_1,\theta)\cdot f(x_2,\theta)\cdot f(x_3,\theta)\cdot\ldots\cdot f(x_n,\theta)]\}'=0$$

基于此求解过程，可以由表格中的汇总数据可得到似然函数：

$$\begin{aligned}\mathrm{L}(\lambda)&=\mathrm{p}^{\mathrm{n}_1}(0<\mathrm{x}<30)\cdot\mathrm{p}^{\mathrm{n}_2}(30<\mathrm{x}<60)\cdot\mathrm{p}^{\mathrm{n}_3}(60<\mathrm{x}<90)\cdot\mathrm{p}^{\mathrm{n}_4}(90<\mathrm{x})\\&=[\mathrm{S}(0)-\mathrm{S}(30)]^{\mathrm{n}_1}[\mathrm{S}(30)-\mathrm{S}(60)]^{\mathrm{n}_2}[\mathrm{S}(60)-\mathrm{S}(90)]^{\mathrm{n}_1}[\mathrm{S}(90)]^{\mathrm{n}_4}\\&=(1-\mathrm{e}^{-30\lambda})^{622}(\mathrm{e}^{-30\lambda}-\mathrm{e}^{-60\lambda})^{558}(\mathrm{e}^{-60\lambda}-\mathrm{e}^{-90\lambda})^{504}(\mathrm{e}^{-90\lambda})^{8316}\end{aligned}$$

其中，$\mathrm{n_1}=622,\ \mathrm{n_2}=558,\ \mathrm{n_3}=504,\ \mathrm{n_4}=8316,$

$\mathrm{p(a<x<b)=S(a)-S(b)}$为分布函数在区间(a,b)的概率，$\mathrm{S(x)}=\mathrm{e}^{-\lambda x}$ 为指数分布的生存函数。

对似然函数取对数可得到对数似然函数：

$$\mathrm{l}(\lambda)=\ln[\mathrm{L}(\lambda)]=622\ln\left(1-\mathrm{e}^{-30\lambda}\right)+558\ln\left(\mathrm{e}^{-30\lambda}-\mathrm{e}^{-60\lambda}\right)$$

$$+504\ln\left(\mathrm{e}^{-60\lambda}-\mathrm{e}^{-90\lambda}\right)+8316(\mathrm{e}^{-90\lambda})$$

通过 R 软件可以画出对数似然函数图像如下：

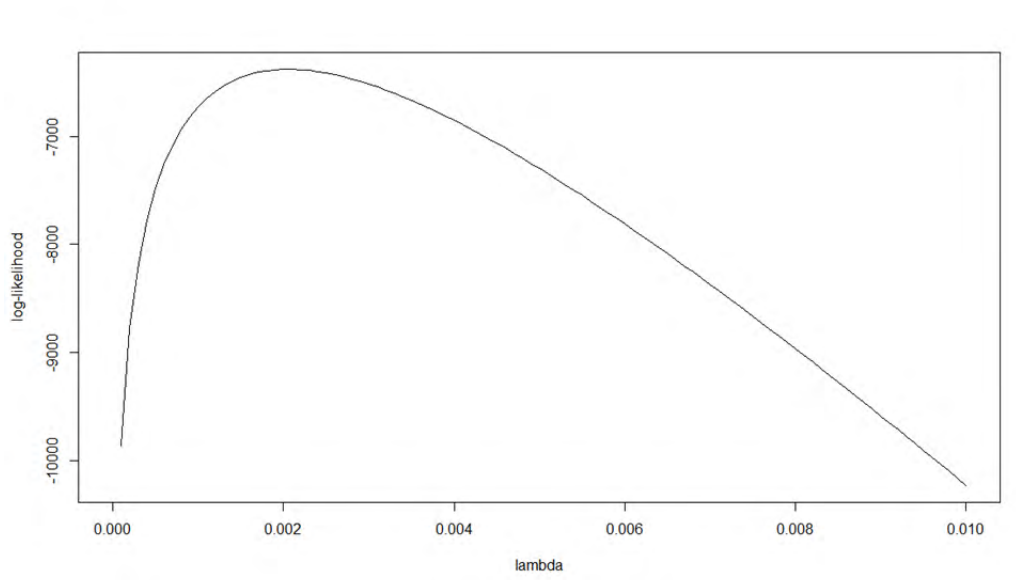

图 1　右截断缺失数据下的对数似然函数图像

直接对对数似然函数的参数 λ 求导得到似然方程：

$$l'(\lambda) =$$

$$\left(622 \cdot 30e^{-30\lambda}\right)/\left(1-e^{-30\lambda}\right) + \left[558 \cdot \left(-30e^{-30\lambda} + 60e^{-60\lambda}\right)\right]/\left(e^{-30\lambda} - e^{-60\lambda}\right) +$$

$$\left[504 \cdot \left(-60e^{-60\lambda} + 90e^{-90\lambda}\right)\right]/(e^{-60\lambda} - e^{-90\lambda}) + [8316 \cdot (-90e^{-90\lambda})]/(e^{-90\lambda})$$

由于似然方程不存在解析解，所以用 R 软件求出 $l'(\lambda)=0$ 的解，求得 λ 的最大似然估计为：$\hat{\lambda}=0.002041668$，失效天数为 $1/0.002041668=487$ 天。

在得到参数 λ 的最大似然估计后，继续求最大似然估计的方差。可以用 Fisher 信息量进行近似，其中 Fisher 信息量的定义如下：

$$I(\lambda) = -E[\partial\, l(\lambda)]$$

$l(\lambda)$是之前求出的对数似然函数，$I(\lambda)$我们称之为费希尔信息量。

最大似然估计的方差我们可以采用费希尔信息量来近似。即：

$$1/I(\lambda) \approx \mathrm{Var}(\hat{\lambda})$$

之前我们已经求出对数似然函数的一阶导数，即：

$$l'(\lambda) = (622 \times 30e^{-30\lambda})/(1-e^{-30\lambda}) +$$

$$[558 \times (-30e^{-30\lambda} + 60e^{-60\lambda})]/(e^{-30\lambda} - e^{-60\lambda}) +$$

$$[504\times(-60e^{-60\lambda}+90e^{-90\lambda})]/(e^{-60\lambda}-e^{-90\lambda})\ +\ [8316\times(-90e^{-90\lambda})]/e^{-90\lambda}$$

对其参数λ继续求导数，得到二阶导数：

$$l''(\lambda)=(-90\times622\times e^{-30\lambda}/(1-e^{-30\lambda})^2\ -\ (90\times558\times e^{-90\lambda})/(e^{-30\lambda}-e^{-60\lambda})^2$$

$$-\quad(504\times360\times e^{-150\lambda})/(e^{-60\lambda}-e^{-90\lambda})^2$$

将λ的最大似然估计值$\hat{\lambda}$=0.002041668代入到上式中，可求出I(λ)=76782348.8，所以$\widehat{Var(\hat{\lambda})}=1/I(\hat{\lambda})$ =0.000000013。

3.1.2 Pareto分布的建模与求解

用R软件产生Pareto分布随机数，其中参数α=3，θ=1000，均值$E(X)=\frac{\theta}{\alpha-1}$ = 500，$Var(X)=(\frac{\theta}{\alpha-1})^2(\frac{\alpha}{\alpha-2})$ = 750000。我们得到了以下区间的汇总数据。

表 2　Pareto分布右截断的汇总数据

天数	(0，30)	(30，60)	(60，90)	90^+
抗体失效的实验体数	842	764	644	7750

由于Pareto分布有两个参数，我们可以固定其中一个参数α =3，然后用最大似然估计法估计另一个参数θ。具体做法如下：

由汇总数据列出似然函数为：

$$L(\theta)=p^{n_1}(0<x<30)\cdot p^{n_2}(30<x<60)\cdot p^{n_3}(60<x<90)\cdot p^{n_4}(90<x)$$

$$=[S(0)-S(30)]^{n_1}[S(30)-S(60)]^{n_2}[S(60)-S(90)]^{n_3}[S(90)]^{n_4}$$

$$=[1-(\frac{\theta}{30+\theta})^3]^{n_1}[(\frac{\theta}{30+\theta})^3$$

$$-(\frac{\theta}{60+\theta})^3]^{n_2}[(\frac{\theta}{60+\theta})^3\ -(\frac{\theta}{90+\theta})^3]^{n_3}[(\frac{\theta}{90+\theta})^3]^{n_4}$$

$$=[1-(\frac{\theta}{30+\theta})^3]^{842}[(\frac{\theta}{30+\theta})^3$$

$$-(\frac{\theta}{60+\theta})^3]^{764}[(\frac{\theta}{60+\theta})^3\ -(\frac{\theta}{90+\theta})^3]^{644}[(\frac{\theta}{90+\theta})^3]^{7750}$$

其中，n_1=842，n_2=764，n_3=644，n_4=7750，p(a<x<b) = S(a) - S(b) 为分布函数在区间（a,b）的概率，$S(x) = (\frac{\theta}{x+\theta})^3$ 为 Pareto 分布的生存函数。

对似然函数取对数可得到对数似然函数：

$$l(\theta) = 842\ln[1-(\frac{\theta}{30+\theta})^3] + 764\ln[(\frac{\theta}{30+\theta})^3-(\frac{\theta}{60+\theta})^3] + 644\ln[(\frac{\theta}{60+\theta})^3-(\frac{\theta}{90+\theta})^3] + 23250\ln(\frac{\theta}{90+\theta})$$

通过 R 软件可以画出对数似然函数图像如下：

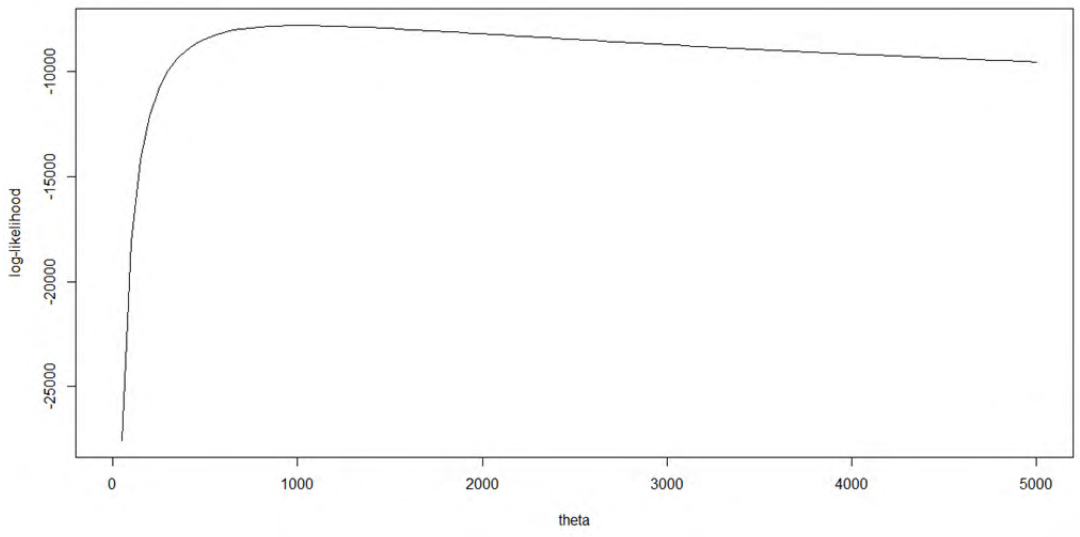

图 2　Pareto 分布右截断缺失数据下的对数似然函数图像

再对参数 θ 求导后，令其为 0 可得到 θ 的最大似然估计，由于方程求导较为复杂，所以直接由 R 软件进行求导得到最大似然估计值为$\hat{\theta}$=1013.4，所以失效天数为$\frac{\hat{\theta}}{\alpha-1}$ = 1013.4/2 ≈ 507 天。

通过 Fisher 信息量近似最大似然估计的方差。将最大似然估计的值代入到对数似然函数的二阶导函数中，可求得其二阶导数值为-0.002，对其取负后再取导数，从而得到最大似然估计的方差为：$Var(\hat{\theta}) = 500$。

3.2 试验二的建模与求解

3.2.1 指数分布的建模求解

我们依旧用 R 软件模拟指数分布的随机数，总体仍为 10000 个实验体,其中参数 λ=0.002，期望值$E(x) = 1/0.002 = 500$天，方差$Var(x) = 1/(500 \times 500) = 0.000004$。其中，（0，100）的数据呈缺失状态，我们得到了（100，130），（130，160），（160，190），190^{+}这四个区间的数据，其汇总数据如下：

表 3　指数分布左缺失的汇总数据

天数	（100，130）	（130，160）	（160，190）	190^{+}
抗体失效的实验体数	471	488	445	6806

我们已知的是：$(x|x>100) \sim f(x)/[1-F(100)]$，其中，x 为变量天数，$F(100)$为指数分布的分布函数在横坐标为 100 处的取值。

所以我们得到似然函数：

$$\begin{aligned} L(\lambda) &= p^{n_1}(100 < x < 130) \cdot p^{n_2}(130 < x < 160) \cdot p^{n_3}(160 < x \\ &< 190) \cdot p^{n_4}(190 < x) \\ &= [S(100) - S(130)]^{n_1} \cdot [S(130) - S(160)]^{n_2} \cdot [S(160) - S(90)]^{n_3} \\ &\cdot [S(190)]^{n_4} / [1 - F(100)]^{n_1+n_2+n_3+n_4} \\ &= (e^{-100\lambda} - e^{-130\lambda})^{471}(e^{-130\lambda} - e^{-160\lambda})^{488}(e^{-160\lambda} \\ &- e^{-190\lambda})^{445}(e^{-190\lambda})^{6806} / (e^{-100\lambda})^{8210} \end{aligned}$$

其中，$n_1 = 471$，$n_2 = 488$，$n_3 = 445$，$n_4 = 6806$，p(a<x<b) = S(a) - S(b)为分布函数在区间(a,b)的概率，$S(x) = 1 - F(x) = e^{-\lambda x}$为指数分布的生存函数。

对似然函数取对数可得到对数似然函数：

$$\begin{aligned} l(\lambda) &= ln[L(\lambda)] \\ &= 471\ln(e^{-100\lambda} - e^{-130\lambda}) + 488\ln(e^{-130\lambda} - e^{-160\lambda}) \\ &+ 445\ln(e^{-160\lambda} - e^{-190\lambda}) + 6806\ln(e^{-190\lambda}) - 8210\ln(e^{-100\lambda}) \end{aligned}$$

通过 R 软件可以画出对数似然函数图像如下：

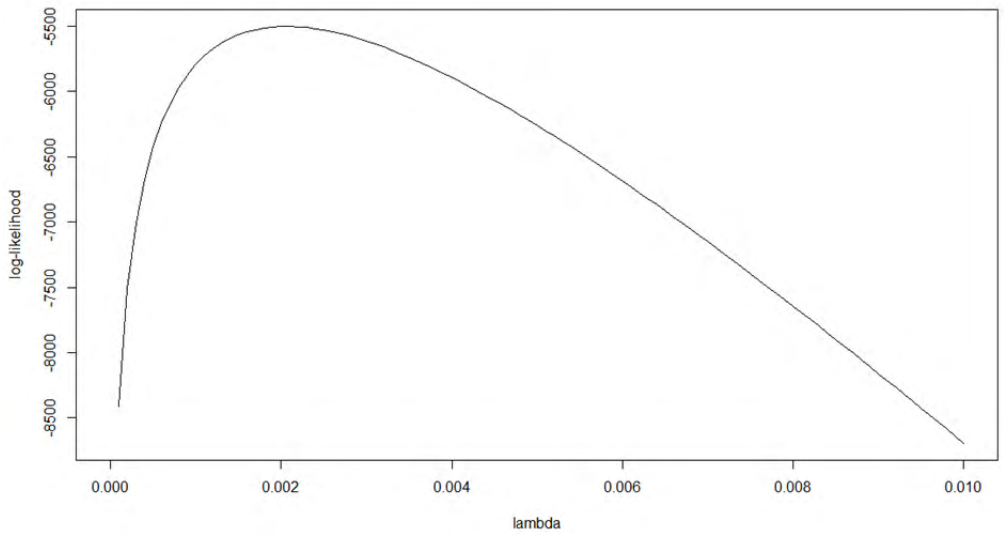

图 3　左缺失数据下的对数似然函数图像

继续对对数似然函数的参数 λ 求导得到似然方程：

$$l'(\lambda) = [471 \times (-100e^{-100\lambda} + 130e^{-130\lambda})]/(e^{-100\lambda} - e^{-130\lambda}) +$$

$$[488 \times (-130e^{-130\lambda} + 160e^{-160\lambda})]/(e^{-130\lambda} - e^{-160\lambda}) +$$

$$[445 \times (-160e^{-160\lambda} + 190e^{-190\lambda})]/(e^{-160\lambda} - e^{-190\lambda}) - 6806 \times 190 + 8120 \times 100$$

该似然方程仍不存在解析解，所以用 R 软件求出 $l'(\lambda)=0$ 的解，求得 λ 的最大似然估计为：$\hat{\lambda}$=0.002047243 ，失效天数为 1/0.002047243 =485 天。

在得到参数 λ 的最大似然估计后 ，再求最大似然估计的方差, 仍旧用 Fisher 信息量的倒数来近似最大似然估计的方差。

首先求得对数似然函数的二阶导数为:

$$l''(\lambda) = -423900e^{-230\lambda}/(e^{-100\lambda} - e^{-130\lambda})^2 - 439200e^{-290\lambda}/(e^{-130\lambda} - e^{-160\lambda})^2$$

$$- 400500e^{-350\lambda}/(e^{-160\lambda} - e^{-190\lambda})^2$$

将 λ 的最大似然估计值$\hat{\lambda}$=0.002047243 代入到上式中，可求出 I(λ)= -334882001，所以$\widehat{Var(\hat{\lambda})}$=1/I($\hat{\lambda}$)= 0.00000000298。

3.2.2 Pareto 分布的建模与求解

我们用 R 软件模拟 Pareto 分布的随机数，总体还是 10000 个实验体，其中参数 $\alpha = 3$，θ =1000，均值 E(X) = $\theta/(\alpha-1)$ = 500，Var(X) = $(\theta/\alpha-1)^2(\alpha/\alpha-2)$ = 750000。我们得到了以下区间的汇总数据。其中，（0，100）的数据呈缺失状态，我们得到了（100，130），（130，160），（160，190），190^+这四个区间的数据，其汇总数据如下：

表 4　Pareto 分布左缺失的汇总数据

天数	（100，130）	（130，160）	（160，190）	190^+
抗体失效的实验体数	591	530	479	5923

我们得到的是条件随机变量：(x|x>100) ~ $f(x)/[1-F(100)]$，其中，x 为天数，f(x)为指数分布的密度函数，F(100)为指数分布的分布函数在 100 处的取值。

所以我们得到似然函数：

$$L(\theta) = p^{n_1}(100 < x < 130) \cdot p^{n_2}(130 < x < 160) \cdot p^{n_3}(160 < x < 190) \cdot p^{n_4}(190 < x)/[1-F(100)]^{n_1+n_2+n_3+n_4}$$

$$= [S(100)-S(130)]^{n_1} \cdot [S(130)-S(160)]^{n_2} \cdot [S(160)-S(190)]^{n_3} \cdot [S(190)]^{n_4}/[1-F(100)]^{n_1+n_2+n_3+n_4}$$

$$= \left[\left(\frac{\theta}{100+\theta}\right)^3 - \left(\frac{\theta}{130+\theta}\right)^3\right]^{n_1} \cdot \left[\left(\frac{\theta}{130+\theta}\right)^3 - \left(\frac{\theta}{160+\theta}\right)^3\right]^{n_2} \cdot \left[\left(\frac{\theta}{160+\theta}\right)^3 - \left(\frac{\theta}{190+\theta}\right)^3\right]^{n_3} \cdot \left[\left(\frac{\theta}{190+\theta}\right)^3\right]^{n_4} / \left(\frac{\theta}{100+\theta}\right)^{n_1+n_2+n_3+n_4}$$

$$= \left[\left(\frac{\theta}{100+\theta}\right)^3 - \left(\frac{\theta}{130+\theta}\right)^3\right]^{591} \cdot \left[\left(\frac{\theta}{130+\theta}\right)^3 - \left(\frac{\theta}{160+\theta}\right)^3\right]^{530} \cdot \left[\left(\frac{\theta}{160+\theta}\right)^3 - \left(\frac{\theta}{190+\theta}\right)^3\right]^{479} \cdot \left[\left(\frac{\theta}{190+\theta}\right)^3\right]^{5923} / \left(\frac{\theta}{100+\theta}\right)^{7523}$$

其中，n_1= 591，n_2= 530，n_3= 479，n_4= 5923，p(a<x<b) = S(a) - S(b)为分布函数在区间(a,b)的概率，$S(x) = 1-F(x) = (\frac{\theta}{x+\theta})^3$ 为 Pareto 分布的生存函数。

对数似然函数如下所示：

$l(\theta) = 591\ln[(\frac{\theta}{100+\theta})^3-(\frac{\theta}{130+\theta})^3] + 530\ln[(\frac{\theta}{130+\theta})^3-(\frac{\theta}{160+\theta})^3] + 479\ln[(\frac{\theta}{160+\theta})^3-(\frac{\theta}{190+\theta})^3] + 5923\ln(\frac{\theta}{190+\theta})-7523\ln(\frac{\theta}{100+\theta})$

通过 R 软件可以画出对数似然函数图像如下：

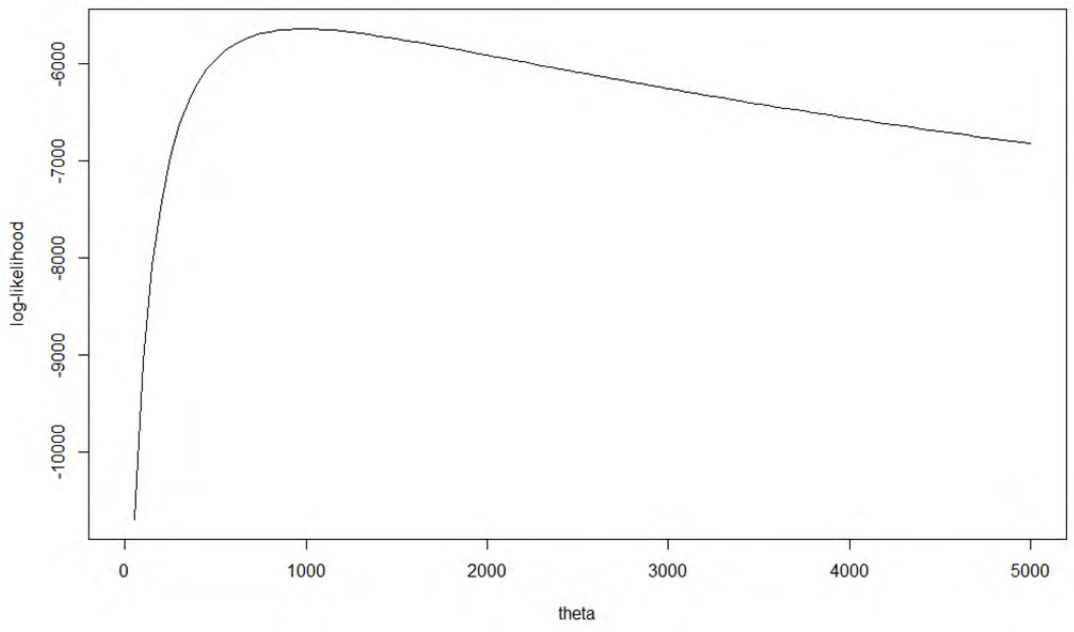

图 4　Pareto 分布左缺失数据下的对数似然函数图像

再对参数 θ 求导后令其为 0 可得到 θ 的最大似然估计为$\hat{\theta}$=984.8，所以失效天数为$\frac{\hat{\theta}}{\alpha-1}$ = 984.8 /2 ≈ 492 天。

最后通过费希尔信息量近似最大似然估计的方差。将最大似然估计的值代入到对数似然函数的二阶导函数中，二阶导函数的求解由 R 软件来实现，所以可求得其二阶导数值为-0.001，对其取负后再取导数，从而得到最大似然估计的方差为：$Var(\hat{\theta}) = 1000$。

综上所述，在带右截断的缺失数据时，指数分布的最大似然估计值为$\hat{\lambda}$=0.002041668<0.002，并且其方差为$\widehat{Var(\hat{\lambda})}=1/I(\hat{\lambda})$ =0.000000013<0.000004。

Pareto 分布的最大似然估计值$\hat{\theta}$=1013.4>1000，其方差为$Var(\hat{\theta}) = 500$<750000。带左缺失数据时，指数分布的最大似然估值$\hat{\lambda}$=0.002047243<0.002，其方差为$\widehat{Var(\hat{\lambda})} = 1/I(\hat{\lambda}) = 0.00000000298$<$0.00004$。Pareto 分布的最大似然估计值$\hat{\theta}$=984.8<1000，相应的方差为$Var(\hat{\theta}) = 1000$<75000。

综合来看，经过与真实值的对比，可以发现 Pareto 分布与真实分布的拟合程度较好。

四、模型的假设检验

假设检验是先对总体参数提出一个假设值, 然后利用样本信息判断这一假设是否成立。我们用假设检验来判断原假设的总体和现在实际的总体是否发生了显著差异。假设检验的一般步骤为：

1. 提出原始假设和替代方案。被检验的断言称为以 H0 表示的主要假设, 与主要假设相反的断言称为以 H1 表示的备择假设。

2. 确定显著性水平, 一般选择区间内估计总体参数的概率误差概率为 0.05。

3. 构造检验统计量。可以用它来确定是否要否定原假设。

4. 根据显著性水平确定拒绝域临界值。所谓的拒绝域是当样本从统计数据中计算出来时, 是在初始假设的基础上划定的区域。代表了偏离最初设定的程度。

5. 计算检验统计量与临界值进行比较，最后得出结论。如果计算出的检验统计量的值落在拒绝域中, 则拒绝原假设, 接受备择假设 ;如果没有落在拒绝域, 就说明拒绝原假设的证据还不够充分。

这里我们可以选用似然比检验，似然比检验和一般的假设检验含义一样，但是效果会更好，都是为了检验模型好坏或是否恰当。似然比是有约束条件下的似然函数最大值与无约束条件下的似然函数最大值之比。因此，似然比检验的实质是比较有约束条件下的似然函数最大值与无约束条件下的似然函数最大值。

4.1 试验一的假设检验

4.1.1 指数分布的似然比检验

考虑总体分布为指数分布，即 X~exp(λ) ,假设设为如下情形。原假设 H_0：λ_0 =0.002 ，备择假设 H_1：λ_1=0.002041668。设显著性水平 α = 0.05，检验统计量选为：$T = 2ln(L_1 - L_0)$ 。

$$L_0 = f(x_1,\lambda_0)\cdot f(x_2,\lambda_0)\cdot f(x_3,\lambda_0)\cdot \ldots \cdot f(x_n,\lambda_0)$$
$$= {\lambda_0}^n \cdot e^{-\lambda_0 x_1}\cdot e^{-\lambda_0 x_2}\cdot \ldots \cdot e^{-\lambda_0 x_n},$$
$$L_1 = f(x_1,\lambda_1)\cdot f(x_2,\lambda_1)\cdot f(x_3,\lambda_1)\cdot \ldots \cdot f(x_n,\lambda_1)$$
$$= {\lambda_1}^n \cdot e^{-\lambda_1 x_1}\cdot e^{-\lambda_1 x_2}\cdot \ldots \cdot e^{-\lambda_1 x_n},$$

其中，n=10000, f(x ,λ)为指数分布的密度函数。

将区间数据代入得到：$L_0 = (1-e^{-30\lambda_0})^{622}(e^{-30\lambda_0}-e^{-60\lambda_0})^{558}(e^{-60\lambda_0}-e^{-90\lambda_0})^{504}(e^{-90\lambda_0})^{8316}$

所以有

$$\ln L_0 = 622\ln(1-e^{-30\lambda_0}) + 558\ln(e^{-30\lambda_0}-e^{-60\lambda_0}) + 504\ln(e^{-60\lambda_0}-e^{-90\lambda_0}) + 8316\ln(e^{-90\lambda_0}) = -6378.891$$

同理有

$$L_1 = (1-e^{-30\lambda_1})^{622}(e^{-30\lambda_1}-e^{-60\lambda_1})^{558}(e^{-60\lambda_1}-e^{-90\lambda_1})^{504}(e^{-90\lambda_1})^{8316}$$
$$\ln L_1 = 622\ln(1-e^{-30\lambda_1}) + 558\ln(e^{-30\lambda_1}-e^{-60\lambda_1}) + 504\ln(e^{-60\lambda_1} - e^{-90\lambda_1}) + 8316\ln(e^{-90\lambda_1}) = -6378.353$$

所以代入数据可知检验统计量$T = 2ln(L_1 - L_0) = 1.076.$

对于似然比检验而言，$[T|H_0$为真$]\sim \chi^2(d)$，其中$\chi^2(d)$是自由度为 d 的卡方分布。本例中，d=1，并且经过查表可知$\chi^2(1)$的 0.05 上侧分位数为 3.841。

因为 1.076<3.841，所以在显著性水平 α=0.05 的条件下接受原假设。

4.1.2 Pareto 分布的似然比检验

考虑总体分布为 Pareto 分布，即 X~Pareto(α，θ) ,其中 α=3，此时假设可设为如下情形。原假设 H_0: θ_0=1000 ，备择假设 H_1: θ_1=1013.4。设显著性水平 α = 0.05，检验统计量选为：$T = 2ln(L_1 - L_0)$。

$$L_0 = f(x_1,\theta_0)\cdot f(x_2,\theta_0)\cdot f(x_3,\theta_0)\cdot \ldots \cdot f(x_n,\theta_0)$$
$$= [1-(\frac{\theta_0}{30+\theta_0})^3]^{842}[(\frac{\theta_0}{30+\theta_0})^3 - (\frac{\theta_0}{60+\theta_0})^3]^{764}[(\frac{\theta_0}{60+\theta_0})^3 - (\frac{\theta_0}{90+\theta_0})^3]^{644}[(\frac{\theta_0}{90+\theta_0})^3]^{7750}$$

$$L_1 = f(x_1,\theta_1)\cdot f(x_2,\theta_1)\cdot f(x_3,\theta_1)\cdot \ldots \cdot f(x_n,\theta_1)$$

$$= [1-(\frac{\theta_1}{30+\theta_1})^3]^{842}[(\frac{\theta_1}{30+\theta_1})^3 - (\frac{\theta_1}{60+\theta_1})^3]^{764}[(\frac{\theta_1}{60+\theta_1})^3 - (\frac{\theta_1}{90+\theta_1})^3]^{644}[(\frac{\theta_1}{90+\theta_1})^3]^{7750}$$

所以$ln L_1 = 842\ln[1-(\frac{\theta_1}{30+\theta_1})^3] + 764\ln[(\frac{\theta_1}{30+\theta_1})^3-(\frac{\theta_1}{60+\theta_1})^3] + 644\ln[(\frac{\theta_1}{60+\theta_1})^3 -(\frac{\theta_1}{90+\theta_1})^3] + 23250\ln(\frac{\theta_1}{90+\theta_1}) = -7790.726$

$ln L_0 = = 842\ln[1-(\frac{\theta_0}{30+\theta_0})^3] + 764\ln[(\frac{\theta_0}{30+\theta_0})^3-(\frac{\theta_0}{60+\theta_0})^3] + 644\ln[(\frac{\theta_0}{60+\theta_0})^3 -(\frac{\theta_0}{90+\theta_0})^3] + 23250\ln(\frac{\theta_0}{90+\theta_0}) = -7790.912$

所以$T = 2ln(L_1 - L_0)$ =0.37<3.841，所以在显著性水平 α=0.05 的条件下接受原假设。

综上所述，在显著性水平 α=0.05 的情况下，指数分布与 Pareto 分布的似然比检验均接受了原假设。

4.2 试验二的假设检验

4.2.1 指数分布的似然比检验

本节仍采用似然比检验，总体分布情况与 4.1.1 无异。此时假设情形如下:原假设为 H_0: λ_0=0.002 ，备择假设为 H_1: λ_1=0.002047243 。并设显著性水平 α = 0.05，$\chi^2(1)$的 0.05 上侧分位数为 3.841。

$$L_0 = [S(100)-S(130)]^{n_1}\cdot[S(130)-S(160)]^{n_2}\cdot[S(160)-S(90)]^{n_3}$$
$$\cdot\ [S(190)]^{n_4}\cdot[1-F(100)]^{n_1+n_2+n_3+n_4}$$
$$= (e^{-100\lambda_0}-e^{-130\lambda_0})^{471}(e^{-130\lambda_0}-e^{-160\lambda_0})^{488}(e^{-160\lambda_0}$$
$$-e^{-190\lambda_0})^{445}(e^{-190\lambda_0})^{6806}/(e^{-100\lambda_0})^{8210}$$

所以$ln L_0 = 471ln(e^{-100\lambda_0}-e^{-130\lambda_0}) + 488ln(e^{-130\lambda_0}-e^{-160\lambda_0}) + 445ln(e^{-160\lambda_0}-e^{-190\lambda_0}) + 6806ln(e^{-190\lambda_0}) - 8210ln(e^{-100\lambda_0}) = -5504.001$

$$L_1 = (e^{-100\lambda_1}-e^{-130\lambda_1})^{471}(e^{-130\lambda_1}-e^{-160\lambda_1})^{488}(e^{-160\lambda_1}$$
$$-e^{-190\lambda_1})^{445}(e^{-190\lambda_1})^{6806}/(e^{-100\lambda_1})^{8210}$$

所以$ln\mathrm{L}_1 = 471ln(\mathrm{e}^{-100\lambda_1} - \mathrm{e}^{-130\lambda_1}) + 488ln(\mathrm{e}^{-130\lambda_1} - \mathrm{e}^{-160\lambda_1}) + 445ln(\mathrm{e}^{-160\lambda_1} - \mathrm{e}^{-190\lambda_1}) + 6806ln(\mathrm{e}^{-190\lambda_0}) - 8210ln(\mathrm{e}^{-100\lambda_0}) = -5503.414$

进而求出检验统计量$T = 2ln(\mathrm{L}_1 - \mathrm{L}_0) = 1.192 < 3.841$。所以在显著性水平α=0.05 的条件下接受原假设。

4.2.2 Pareto 分布的似然比检验

总体分布情况与 4.1.2 无异。此时假设情形如下。原假设 H_0 : θ_0=1000 ，备择假设 H_1 : θ_1=984.8。并设显著性水平 α = 0.05，$\chi^2(1)$的 0.05 上侧分位数为 3.841。

$$\mathrm{L}_0 = [S(100) - S(130)]^{n_1} \cdot [S(130) - S(160)]^{n_2} \cdot [S(160) - S(90)]^{n_3} \cdot [S(190)]^{n_4} \cdot [1 - F(100)]^{n_1+n_2+n_3+n_4}$$

$$= [(\frac{\theta_0}{100+\theta_0})^3 - (\frac{\theta_0}{130+\theta_0})^3]^{591}[(\frac{\theta_0}{130+\theta_0})^3 - (\frac{\theta_0}{160+\theta_0})^3]^{530}[(\frac{\theta_0}{160+\theta_0})^3 - (\frac{\theta_0}{190+\theta_0})^3]^{479} (\frac{\theta_0}{190+\theta_0})^{17769} / (\frac{\theta_0}{100+\theta_0})^{22569}$$

$$ln\mathrm{L}_0 = 591\ln[(\frac{\theta_0}{100+\theta_0})^3 - (\frac{\theta_0}{130+\theta_0})^3] + 530\ln[(\frac{\theta_0}{130+\theta_0})^3 - (\frac{\theta_0}{160+\theta_0})^3] + 479\ln[(\frac{\theta_0}{160+\theta_0})^3 - (\frac{\theta_0}{190+\theta_0})^3] + 17769\ln(\frac{\theta_0}{190+\theta_0}) - 22569\ln(\frac{\theta_0}{100+\theta_0}) = -5645.093$$

$$\mathrm{L}_1 = [(\frac{\theta_1}{100+\theta_1})^3 - (\frac{\theta_1}{130+\theta_1})^3]^{591}[(\frac{\theta_1}{130+\theta_1})^3 - (\frac{\theta_1}{160+\theta_1})^3]^{530}[(\frac{\theta_1}{160+\theta_1})^3 - (\frac{\theta_1}{190+\theta_1})^3]^{479} (\frac{\theta_1}{190+\theta_1})^{17769} / (\frac{\theta_1}{100+\theta_1})^{22569}$$

$$ln\ \mathrm{L}_1 = 591\ln[(\frac{\theta_1}{100+\theta_1})^3 - (\frac{\theta_1}{130+\theta_1})^3] + 530\ln[(\frac{\theta_1}{130+\theta_1})^3 - (\frac{\theta_1}{160+\theta_1})^3] + 479\ln[(\frac{\theta_1}{160+\theta_1})^3 - (\frac{\theta_1}{190+\theta_1})^3] + 17769\ln(\frac{\theta_1}{190+\theta_1}) - 22569\ln(\frac{\theta_1}{100+\theta_1}) = -5644.951$$

进而求出检验统计量$T = 2ln(\mathrm{L}_1 - \mathrm{L}_0) = 0.29 < 3.841$。所以在显著性水平α=0.05 的条件下接受原假设。

综上所述，在显著性水平α=0.05 的情况下，指数分布与 Pareto 分布的似然比检验同样也接受了原假设。

五、结论与建议

本文的总体思路为先通过模拟两种缺失数据，再对两种试验设计分别建立指数分布模型与 Pareto 分布模型，分析各自的最大似然函数值及其方差，并且与真实值相对比，最后用似然比检验验证了模型的可行性。结果表明，疫苗的失效时间用 Pareto 分布模拟较为合理，其拟合误差相对较小。本文采用的方法主要为：有缺失数据时的最大似然估计法，经过与真实值的对比，发现其确实可行，并且与真实数据的拟合程度较好，这也证明了该方法的优良性。

本文主要的创新点有以下两点：

1. 中国大部分疫苗临床实验都在巴西进行的，因为国内几乎没有感染新冠病毒的患者，所以相关患者的实验做不了，数据也拿不到。而且就连一个人的抗体浓度实时变化情况也不知道，因为不可能对一个人进行连续的抽血测试，那他的疫苗失效的时间点就不清楚。但尽管如此，依靠本文的方法，我们依旧可以算出相应的疫苗失效天数，进而知道疫苗有效性能够持续多久。而本文的目的也是希望能够为相关的疫苗有效性研究提供相应的参考，

2. 如今全国范围内在接种新冠疫苗，大部分人都接种了两针。在这样的现实背景下，会产生一个特别庞大的数据量。若有人想抽血检查体内的抗体是否还有效，则只需满足相关试验设计，并且知道总体分布情况，就可以得到疫苗的失效天数。这对研究疫苗有效性有一定的借鉴意义。

本文利用了两个连续型分布模型对疫苗的有效性进行了分析，综合评价本文的研究可以发现仍旧存在一定的不足，主要的不足之处及改进的方向有以下几点：

（1）本文仅采取了指数分布与 Pareto 分布模型，分别对应了单参数和两参数的连续型分布，但理论上有更多连续型分布可供选择，如伽马分布、韦布尔分布、卡方分布等等。在以后的研究中也可以尝试模拟这些分布，关于疫苗失效时间的具体分布也会更加明确。其难点在于某些分布函数没有具体的表达式，需要对其概率密度函数进行积分，这对模型的建立有一定的难度，也是未来研究中需要考虑并改进的地方。

（2）在带有左缺失数据试验设计的部分，刚开始为了模型的简便，我们考虑到现实中的干扰因素较多，所以将第一次测试无效的对象都舍去了。但实际上人们接种的疫苗失效后，政府一定会考虑到给一部分人接种第三针，即加强针，使他们体内的抗体浓度恢复并提高，此时这些实验体的变化会给我们的模型增加一些难度。在可靠性理论有一个“ 修旧如旧 ”的概念，我们可以将这个问题和“ 修旧如旧 ”概念相结合，即接种了加强针后的实验体重新恢复成左缺失数据，这时再考虑模型会不会有改进。

（3）在选用双参数的 Pareto 分布时，我们固定了其中一个变量 α，用最大似然估计法估计了另外一个参数 θ。后续可以考虑两参数都不固定的情形，使模型更加复杂化，这时候在考虑如何用最大似然函数法求解。

参考文献

1. Moderna 2nd Annual Vaccines Day——Predicting the protection of SARS-CoV-2 , Miles P Davenport..

2. Antibody-dependent enhancement and SARS-CoV-2 vaccines and therapies , Wen shi lee, Adam K. Wheatley, Stephen J. Kent, Brandon J. Dekosky.

3. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Nuno R. Faria, Thomas A.Mellan, charles Whittaker, Ingra M.Claro.

4. Maturation and ppersistence of the anti-SARS-CoV-2 memory B cell response , Pascal Chappert

附录

附件 1：模拟右截断缺失数据程序

```
set.seed(1)
```

```
x = rexp(n = 10000, rate = 0.002)
hist(x, breaks = 20, freq = TRUE, labels = TRUE)
length(x[x>0 & x<30])
length(x[x>30 & x<60])
length(x[x>60 & x<90])
length(x[x>90])
library("actuar")
set.seed(3)

x = rpareto(n = 10000, shape = 3, scale = 1000)
length(x[x>0 & x<30])
length(x[x>30 & x<60])
length(x[x>60 & x<90])
length(x[x>90])
```

附件 2：模拟左缺失数据程序

```
set.seed(2)
x = rexp(n = 10000, rate = 0.002)
hist(x, breaks = 20, freq = TRUE, labels = TRUE)
length(x[x<100])
length(x[x>100 & x<130])
length(x[x>130 & x<160])
length(x[x>160 & x<190])
length(x[x>190])
```

```
set.seed(4)
x = rpareto(n = 10000, shape = 3, scale = 1000)
length(x[x<100])
length(x[x>100 & x<130])
length(x[x>130 & x<160])
length(x[x>160 & x<190])
length(x[x>190])
```

附件 3：似然方程求解程序

```
f = function(t){
  y =
622*log(1-exp(-30/t))+558*log(exp(-30/t)-exp(-60/t))+504*log(exp(-60/
t)-exp(-90/t))+8316*(-90/t)
  return(y)
}
optimize(f, c(0, 10000), maximum=TRUE)

f = function(t){
  y =
-8120*(-100/t)+471*log(exp(-100/t)-exp(-120/t))+488*log(exp(-130/t)-e
xp(-160/t))+445*log(exp(-160/t)-exp(-190/t))+6806*(-190/t)
  return(y)
}
```

```
optimize(f, c(0, 10000), maximum=TRUE)
```

附件 4：Fisher 信息量求解程序

```
f = function(t){
  y1 =
(15919800*exp(-260*t)-12669900*exp(-230*t))/(exp(-100*t)-exp(-130*t))
^2
  y2 =
(24985600*exp(-320*t)-20740000*exp(-290*t))/(exp(-130*t)-exp(-160*t))
^2
  y3 =
(32129000*exp(-380*t)-27456500*exp(-350*t))/(exp(-160*t)-exp(-190*t))
^2
  y = y1+y2+y3
  return(y)
}
f1 = expression(842*log(1-(t/(30+t))^3))
f2 = expression(764*log((t/(30+t))^3-(t/(60+t))^3))
f3 = expression(644*log((t/(60+t))^3-(t/(90+t))^3))
f4 = expression(23250*log(t/(90+t)))
df1 = deriv3(f1, "t", function.arg = TRUE)
df2 = deriv3(f2, "t", function.arg = TRUE)
df3 = deriv3(f3, "t", function.arg = TRUE)
df4 = deriv3(f4, "t", function.arg = TRUE)
```

```
df1(1013.429)
df2(1013.429)
df3(1013.429)
df4(1013.429)
0.0007282584 + 0.0004989896 + 0.0002953594 - 0.003542271

f1 = expression(591*log((t/(100+t))^3-(t/(130+t))^3))
f2 = expression(530*log((t/(130+t))^3-(t/(160+t))^3))
f3 = expression(479*log((t/(160+t))^3-(t/(190+t))^3))
f4 = expression(17769*log(t/(190+t))-22569*log(t/(100+t)))
df1 = deriv3(f1, "t", function.arg = TRUE)
df2 = deriv3(f2, "t", function.arg = TRUE)
df3 = deriv3(f3, "t", function.arg = TRUE)
df4 = deriv3(f4, "t", function.arg = TRUE)
df1(1013.429)
df2(1013.429)
df3(1013.429)
df4(1013.429)
0.0001318348 + 3.296533e-05 - 4.149669e-05 - 0.00126186
```

附件 5：最大似然函数图及最大似然函数值求解程序

```
f = function(t){
  y                                                    =
```

```
622*log(1-exp(-30*t))+558*log(exp(-30*t)-exp(-60*t))+504*log(exp(-60*
t)-exp(-90*t))+8316*(-90*t)
  return(y)
}
optimize(f, c(0, 0.1), maximum=TRUE)
plot(f, xlim=c(0, 0.01), xlab="lambda", ylab="log-likelihood")
f(0.002)
f(0.002041668)

f = function(t){
  y                                                                    =
-8120*(-100*t)+471*log(exp(-100*t)-exp(-120*t))+488*log(exp(-130*t)-e
xp(-160*t))+445*log(exp(-160*t)-exp(-190*t))+6806*(-190*t)
  return(y)
}
optimize(f, c(0, 0.1), maximum=TRUE)

plot(f, xlim=c(0, 0.01), xlab="lambda", ylab="log-likelihood")
f(0.002)
f(0.002047243)

f = function(t){
  y                                                                    =
842*log(1-(t/(30+t))^3)+764*log((t/(30+t))^3-(t/(60+t))^3)+644*log((t
```

```
/(60+t))^3-(t/(90+t))^3)+23250*log(t/(90+t))
  return(y)
}
optimize(f, c(0, 2000), maximum=TRUE)

plot(f, xlim=c(0, 5000), xlab="theta", ylab="log-likelihood")

f = function(t){
  y                                                                    =
591*log((t/(100+t))^3-(t/(130+t))^3)+530*log((t/(130+t))^3-(t/(160+t)
)^3)+479*log((t/(160+t))^3-(t/(190+t))^3)+17769*log(t/(190+t))-22569*
log(t/(100+t))
  return(y)
}
optimize(f, c(0, 2000), maximum=TRUE)
plot(f, xlim=c(0, 5000), xlab="theta", ylab="log-likelihood")
```

致谢

值此论文撰写完成之际，我要由衷地感谢我的指导老师 XXX 老师。在此次

论文的选题上，XXX 老师结合了国家的热点话题及自己所擅长的方向，给予了我们很大的帮助。从数据的分析、模拟到梳理框架等等，老师都为我们付出了宝贵的时间。在论文的修改过程中也提供了宝贵的意见。在此由衷的感谢 XXX 老师对我们的无私帮助！

同时，还要感谢我的队员！我们一起相互帮助，相互支持，共同克服一个又一个困难。有了大家的智慧，我们的论文才得以在规定时间内顺利完成。

最后对帮助和关心过我们的老师、同学、长辈给予最真心的感谢！