分类号： G43　　密级： 无
单位代码： 10028　　学号： 2150502002

首都师范大学硕士学位论文

RACUF 累积和控制图在监测长期病患生存时间的应用

研 究 生：袁 钰
指导教师：胡 涛
学科专业：应用统计
研究方向：应用统计

2018 年 2 月 13 日

首都师范大学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

首都师范大学位论文授权使用声明

本人完全了解首都师范大学有关保留、使用学位论文的规定，学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版。有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅。有权将学位论文的内容编入有关数据库进行检索。有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

学位论文作者签名：

日期：　　年　　月　　日

摘　要

生存分析是研究缺失数据的一种数理统计方法。近年来随着计算机技术的发展，生存分析方法在工业、医学、保险、经济学等众多领域展现其强大的分析和预测能力。

统计控制过程中的控制图方法在疾病监测中发挥着重要作用，是生存分析的重要方法之一。本文针对传统控制图存在的不足，考虑长期病患被治愈的情况，优化了处理生存数据在追溯期中发生大比例删失情况的方法。主要的工作如下：

1、介绍传统控制图和累积和控制图的原理，结合 Weibull 分布的特点，利用回归模型拟合不含治愈情况的生存数据，并且使用该方法进行检测，得到风险调整生存时间（Risk Adjusted Survival Time，RAST）累积和控制图。

2、考虑治愈情况的生存数据的特点，使用似然方法，优化 RAST 累积和控制图，得到带有治愈率的风险调整生存时间（Risk Adjusted Survival Time with Cure Fraction，RACUF）累积和控制图。

3、利用模拟数据和实际数据对两种方法进行比较，指出本文的 RACUF 累积和控制图具有较好的应用前景。

关键词: 生存分析；累积和控制图；RAST；RACUF；PT 模型

ABSTRACT

Survival analysis is a mathematical statistics method for the study of missing data. In recent years, with the development of computer technology, survival analysis method has shown strong analytical and predictive ability in many fields such as industry, medicine, insurance, economics and so on.

The control chart in the process of statistical control plays an important role in disease monitoring, and it is one of the most important methods of survival analysis. Aiming at the shortcomings of traditional control charts and considering the healing of long-term patients, this paper optimizes the method of dealing with the large proportion of deleted data in the traceability period. The main work is as follows:

1, Based on the traditional control chart and CUSUM control chart, combined with the features of the Weibull distribution, this paper expands survival data without cure patients by regression model, and use the method of detection, RAST CUSUM control chart.

2, Considering the characteristics of the survival data, this paper uses the likelihood method to optimize the RAST cumulative sum control chart and get the Risk Adjusted Survival Time with Cure Fraction cumulative sum control chart with the cure rate.

3. Comparing the two methods with the simulated data and the actual data, it is pointed out that the RACUF accumulation and control chart in this paper has a good application prospect.

KEY WORDS: Survival analysis; cumulative sum control charts; RAST; RACUF; PT model

目　录

第 1 章　绪论

1.1　研究背景及意义

生存分析是 20 世纪六七十年代发展起来的数理统计学科的新分支，其产生是为要解决生物学、现代医学等大量科学研究中存在着的实际问题。生存分析注重于对生存数据进行统计分析研究，即基于对生物或人类的生存时间的测试和调查取得数据，通过对这些数据的分析和推断来研究生存时间与众多影响因子间关系大小程度的方法 [1−4]。

某些研究，如社会学领域的寻访时间研究及失业和再就业的研究等，虽然与生存本身没有多大的关系，但因为这类研究的数据往往由于某些原因不能被完全观察到，因此需要用特殊的方法对此进行统计处理。这种特殊的方法同样源于对生存资料的统计，因此它也可以被归类到生存分析研究领域。例如，医学领域的某一疾病的发病时间以及初次治疗后的复发时间，机械及系统的失效无法工作的时间，发行货币基金的违约时间等。到目前为止，生存分析已经成为了一种成熟的分析方法，广泛应用于医学、生物学、保险精算学、可靠性工程学、公共卫生学、经济学、金融以及人口统计学等领域 [5]。同时生存分析不仅应用广泛，而且其研究方法十分丰富。特别是 20 世纪九十年代以来，随着计算机技术的发展，高速计算和大规模数据存储使得十几年前不能实现的方法如今可以有效快速地实施。

统计控制过程（Statistics Process Control,SPC）是产品质量与设计中的重要研究内容，它包含一些用来降低产品质量波动以使产品质量保持稳定的诸多有效工具。其中一个现今应用最为广泛的工具就是控制图（control chart）。自 1925 年 Shewhart 博士提出第一个控制图 Shewhart$\bar{X}$，到目前为止，关于控制图的研究成果和应用成果已经相当丰富，并取得了很好的社会效益和经济效益 [6]。

如果将 SPC 看作长期病人的生存情况，影响工程的因素在后者成为影响其生存时间的因素，那么控制图方法就可以在生存分析领域得以利用。但两者仍有区别，如生存分析中有患者被治愈的情况，这会直接影响观测的结果，而这在 SPC 中是不存在的。传统的控制图方法并不考虑治愈率，导致监测病人生存时间的控制图不够灵敏，结果也有一定的误差。研究人员应当重视长期病患被治愈的情况，控制监测误差，提高监测的精度，这也是本文的研究意义所在。

1.2　研究方法及研究现状

统计控制过程（SPC）广泛应用于工业指标的监测和控制，由一些统计工具组成。然而 SPC 也可应用于疾病监测，特别是控制图 [7]。而与工业中控制图的应用不同的是，生存分析中的变量，如性别、年龄、过往病史、术前风险等等难以保证其同质性，因此控制图在生存分析，特别是应用于监测长期病人时，必须将每个病人的风险情况考虑在内，不断加以调整，以期起到准确监测患者情况，有效反馈的效果。

1954 年 Page 提出的累积和控制图（CUSUM chart）较好地监测到了工业生产中微小且连续的变化 [8]。后来的 Grigg 和 Farewell 结合生存数据的特点，对整个生存分析中的控

制图应用进行概述，并且提出了主要用于监测二分问题的风险调整累积和控制图（RAST CUSUM chart）[9]。Steiner 等考虑治疗是否成功的二分变量对控制图进行了改进 [10]。这些研究人员的共同点是监测的患者均为经治疗存活超过 30 天以上，并且都使用了基于 logistic 回归模型的似然得分对这些患者的生存时间进行监测。Biswas 和 Kalbfleisch 提出用 Cox 模型分析死亡时间，并得到平均运行长度（Average Run Length, ARL）的近似值 [11]。Gandy 根据失控状态和受控状态的风险率偏似然比提出了更为普遍的累积和控制图 [12]。Sego 等使用加速死亡模型进行生存数据的研究，这种控制图就是推广的风险调整累积和控制图（RAST CUSUM chart），这种控制图的评分用 Weibull 和 log-logistic 两种回归模型得到 [13]。这些方法在模拟研究中较之于早期的方法监测出异常的速度更快。Sun 和 Kalbfleisch 关注比预期更短或更长的生存时间，提出了可以给出范围的风险调整期望累积和控制图（risk-adjusted Observed-Expected CUSUM chart）[14].Phinikettos 和 Gandy 研究了监测生存数据非参数方法 [15]。Sun、Kalbfleisch 和 Schaubel 提出了用于监测删失了的生存数据的加权累积和控制图 [16]，张敏等研究了 RAST 累积和控制图的预测误差控制 [17]。

在上述研究中，生存数据可能右删失的情况更为普遍，研究人员往往采取条件生存模型，但是往往忽略患者被治愈的情况，假定每一个患者都可能在治疗后死亡，尽管某些患者的生存时间观测不到（删失）[18]。因此自然的想法就是基于治愈患者的比例（治愈率）对控制图进行扩展和和补充，这是本文尝试进行的工作。

1.3 论文的主要工作

本文针对传统控制图存在的不足，引入了一种基于治愈率的累积和控制图，即 RACFU 累积和控制图（Risk-Adjusted with Cure Fraction Cumulative Sum Chart）。利用该方法对生存数据的追溯期中发生的较大比例的删失情况进行分析。本文的研究工作包括以下几个方面：

1. 较为清晰地介绍了传统的控制图方法的原理，以及如何绘制累积和控制图，如何使用控制图监测生存数据的异常情况；

2. 结合生存时间服从 Weibull 分布的特点，利用 AFT 模型拟合不含治愈情况的生存数据，并且使用 RAST 方法进行监测，得到 RAST 累积和控制图；

3. 分析有治愈情况的生存数据，并使用似然方法，采用 PT 模型拟合该数据，再用 RACUF 模型处理，并且模拟研究两种方法进行比较；

4. 使用 RAST 和 RACUF 方法研究一组实际数据，指出该方法应用的广阔前景。。

1.4 论文的组织结构

本文的组织结构如图 1.1 所示。

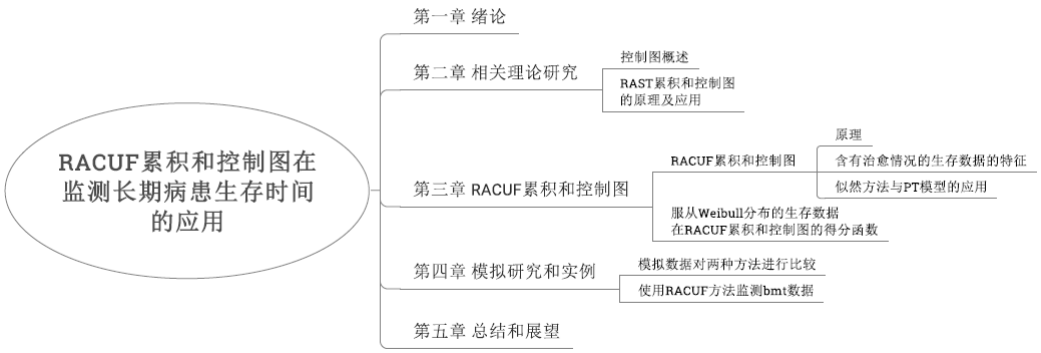

图 1.1　论文的组织结构图

本文主要对含有治愈情况的生存数据进行研究，各章节内容安排如下：

第一章对整篇论文的研究工作作了简要的介绍，阐述生存分析的含义和发展现状，原本用于工业研究的控制图方法在分析生存数据的应用，建立了文章的结构，说明了课题的主要工作。

第二章对控制图进行概述，介绍传统的控制图和累积和控制图，给出绘制累积和控制图的得分函数，阐述累积和控制图的工作原理，并且给出基于 AFT 模型的 RAST 累积和控制图的得分函数。

第三章介绍了考虑长期病患的监测中出现的治愈情况，而提出的 RACUF 累积和控制图，较为详细地给出 RACUF 方法的原理和推导过程，给出含治愈情况的生存数据的似然函数，并且引入 PT 模型，得到服从 Weibull 分布的生存数据在 RACUF 累积和控制图的得分函数。

第四章对前文的方法进行模拟研究和实际应用。随机生成服从 Weibull 分布的数据，重复试验 1000 次验证 RAST 和 RACUF 两种方法的有效性，并且进行比较。使用这两种方法分析 1987 年 Kersey 等研究急性淋巴细胞白血病所使用的数据（bmt）[19]，验证了前文的结论。

第五章总结了本文所做的研究工作，之后对 RACUF 方法的进一步发展做出了展望。

第 2 章　相关理论研究

2.1　控制图概述

2.1.1　传统控制图

控制图方法广泛应用于监测工业过程的稳定性 [20]，世界上第一个控制图 Shewhart$\bar{X}$ 控制图由 Shewhart 博士与 1925 年提出，研究表明它对监测较大的漂移（shift）和异常点 (outlier) 效果明显，并且由于其操作简单，在现在的生产过程中仍有较为广泛的应用。经过几十年的发展，用于检测中小漂移的有效控制图也被提出，累积和控制图（CUSUM chart）就属于这一类控制图。

2.1.2　累积和控制图

累积和控制图是 Page 于 1954 年提出的，这种方法基于 Wald 检验方法，将根据每个观测样本的情况给出该样本权重（得分）[21]。如此累积和控制图不是孤立地使用每个观测提供的信息，而是将整个过程看成一个整体，将所有观测的信息累积起来，从而得出最终的结论。

2.1.3　累积和控制图得分函数

累积和控制图是由每个观测的得分情况绘制而成。设 $i = 1, 2, \ldots$ 表示观测患者的编号，$L(\xi|D_i)$ 是第 i 个患者生存时间的似然函数，其中 ξ 表示似然函数的参数向量，D_i 表示第 i 个病人的生存情况，而病人的生存风险则用一个回归模型计量。

在一个受控状态下，参数向量 ξ 的期望值 ξ_0 被认为是 ξ 的真实值，而 ξ_0 是由受控状态下的历史数据估计的。本文采用通常的假定，即受控数据的模型已知，并且估计误差极小可以忽略 [22]。

下面给出对数似然比形式的累积和控制图的得分函数 W_i:

$$W_i = log\left[\frac{L(\xi_1|D_i)}{L(\xi_0|D_i)}\right] = l(\xi_1|D_i) - l(\xi_0|D_i) \tag{2.1}$$

其中 ξ_1 是失控状态下似然函数的参数向量，$l(.)$ 是对数似然函数。受控状态下的 $\xi = \xi_0$ 和失控状态的 $\xi = \xi_1$ 的改变说明过程发生了本质的变化。

2.1.4　累积和控制图的工作原理

累积和控制图就是绘制得分函数的图像：

$$Z_i = max(0, Z_{i-1} + W_i), i = 1, 2, \ldots \tag{2.2}$$

其中 $Z_0 = 0$。当 $Z_i > h$ 时控制图发出警报，h 就是该控制图的控制线。经研究，通过控制线监测数据发生较大变动的方法是渐进最优的 [23]。

公式 1 中 W_i 越大，累积和控制图中统计量 Z_i 的累积误差就越大，最终超过控制线的上界，表明该过程已经从受控状态到失控。另外累积和控制图只考虑得分为正的情况。

2.2 RAST 累积和控制图

下面介绍 Sego 等提出的风险调整累积和控制图（RAST CUSUM chart），此控制图是基于加速死亡模型（Accelerated Failure Time，AFT）构造的。

2.2.1 AFT 模型

令 T_i 表示第 i 个病人发生事件的时刻，$\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^T$ 为一给定的协变量向量，则 AFT 模型可表示如下：

$$logT_i = \mu + \boldsymbol{\gamma^T x_i} + \sigma V_i, i = 1, \ldots, n \tag{2.3}$$

其中 V_i 是彼此独立同分布的随机误差，且其分布与 $\boldsymbol{x_i}$ 独立。向量 $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$，$\mu$ 和 σ 是未知的参数。

AFT 模型假定协变量向量为 $\boldsymbol{x_i}$ 的第 i 个病人在给定时刻的生存函数与基准生存函数相同。此假定保证了 AFT 模型取对数后易于估计未知参数 μ 和 σ，并且也有较多的分布可以用来拟合 AFT 模型中生存时间的分布情况。考虑到灵活性和可证明的拟合生存时间的优势，本文使用威布尔分布（Weibull Distribution）拟合 AFT 模型中生存时间的情况。

2.2.2 RAST 模型的得分函数

生存时间右删失的情况较为普遍，本文假定删失情况均为右删失。设 C_i 表示相互独立的删失时刻，$i = 1, \ldots, n$。假定删失时刻 C_i 与第 i 个病人发生事件的时刻 T_i 独立，C_i 的分布不依赖未知参数。再设 $\delta_i = 1$ 表示非删失，$\delta_i = 0$ 表示删失，令 $Y_i = min(T_i, C_i)$。这样个体的信息可由随机变量对 (Y_i, δ_i) 和协变量向量与 $\boldsymbol{x_i^T}(i = 1, \ldots, n)$ 表示。

若将未知参数用向量 $\boldsymbol{\eta}$ 表示，那么第 i 个个体生存时间的对数似然函数可表示为

$$log[f(t_i|\boldsymbol{\eta})^{\delta_i} S(t_i|\boldsymbol{\eta})^{1-\delta_i}] \tag{2.4}$$

其中 $S(t_i|\boldsymbol{\eta})$ 表示给定协变量向量 $\boldsymbol{x_i}$，T_i 的生存函数族，$f(t|\boldsymbol{\eta}) = -\frac{\partial}{\partial t} S(t_i|\boldsymbol{\eta})$ 为相应的概率密度函数。

当用 Weibull 分布拟合个体的生存时间的时候，若忽略协变量影响，为保证参数估计的简洁，设 $T_i \sim Weibull(\alpha, \lambda exp(\boldsymbol{\gamma^T} x_i))$。其中 $\alpha = \frac{1}{\sigma}$，$\lambda = e^{\mu}$ 是 T_i 所服从的 Weibull 分布的

两个参数，$\boldsymbol{\gamma}$ 是回归系数向量，V_i 是彼此独立同分布的随机误差。

那么就得到了 Weibull AFT 模型的生存函数：

$$S(t_i|\boldsymbol{\eta}) = exp\{-(\frac{t_i}{\lambda exp(\boldsymbol{\gamma^T x_i})})^{\alpha}\} \tag{2.5}$$

其中 $\boldsymbol{\eta} = (\alpha, \lambda, \boldsymbol{\gamma}^T)^T$。一般假定过程受控时，$\lambda = \lambda_0$。检测过程是否失控，则是观测到新个体时，检测 $\lambda_1 = \rho_1 \lambda_0$。相应的 $\boldsymbol{\eta_0} = (\alpha, \lambda_0, \boldsymbol{\gamma}^T)^T$ 可由受控状态下的训练数据得到其估计值，并且由于受控状态下假定模型是准确的，那么参数估计时的误差可以忽略。

易证明当平均生存时间缩短时，参数 λ 取值范围为 $(0, 1)$；反之有 $\lambda > 1$。Sego 给出了对数似然比形式的 Weibull 分布累积和控制图的得分函数 W_i [24]：

$$W_i = (1 - \rho_1^{-\alpha})(\frac{t_i}{\lambda_0 exp(\boldsymbol{\gamma}^T \boldsymbol{x_i})})^{\alpha} - \delta_i \alpha log \rho_1 \tag{2.6}$$

可从累积和控制图的图像中看出参数 λ 的变化。

第 3 章　RACUF 累积和控制图

本节使用 Rodrigues 等提出的统一治愈率模型（Unified Cure Rate Model）推导累积和控制图的得分函数 [25]。

3.1　RACUF 累积和控制图概述

设总体中的一个样本 i，其是否发生我们感兴趣的事件受 M_i 个因素的影响，对应的概率为 $p_\theta(m_i) = P_\theta(M_i = m_i)$。随机变量 $R_{i1}, R_{i2}, \ldots, R_{im_i}$ 表示第 j 个因素 $(j = 1, 2, \ldots, M_i)$ 使得第 i 个个体发生事件的时刻，彼此独立且同分布，它们的分布函数 $F(z|\boldsymbol{\eta})$，生存函数 $S(z|\boldsymbol{\eta}) = 1 - F(z|\boldsymbol{\eta})$，$\boldsymbol{\eta}$ 是参数向量。

再设随机变量 T_i 表示时间发生的等待时间：$T_i = min(R_{i0}, R_{i1}, R_{i2}, \ldots, R_{im_i})$，其中 $P(R_{i0} = \infty) = 1$，如此就保证了对疾病免疫的个体有无限的生存时间，而 $M_i = 0$ 说明该事件没有发生的可能性。

3.1.1　含有治愈情况的生存数据

下面进一步解释潜变量 M_i 和 R_i：设个体 i 患有某种癌症，经过首次治疗后仍然存活，具有转移性癌细胞的数目为 M_i。当 $M_i = 0$ 时这位患者被治愈，而当 $M_i = m$ 时，该患者有 m 个转移性癌细胞，随机变量 $R_{i0}, R_{i1}, R_{i2}, \ldots, R_{im_i}$ 分别是这 m 个癌细胞可能导致个体 i 癌症再次发病的时刻。如此，引入潜变量 M_i 和 R_i，就将患者可能被治愈的情况考虑在内，并且易于量化未被治愈的患者发病的情况 [26]。

3.1.2　长期病人的生存时间函数

因此，含有治愈情况的长期患者生存时间 T_i 的函数可以用下式表达：

$$S_p(t) = P(T_i > t) = P_\theta(0) + \sum_{m_i=1}^{\infty} p_\theta(m_i) S(t|\boldsymbol{\eta})^{m_i} \tag{3.1}$$

一般的生存函数 $S(t|\boldsymbol{\eta})$ 应当满足 $\lim_{t\to\infty} S(t|\boldsymbol{\eta}) = 0$，表示治疗后的存活时间趋于无限时，患者没有被治愈的情况。而上式中有 $\lim_{t\to\infty} S_p(t) = p_\theta(0) > 0$。长期患者生存时间 T_i 的概率密度函数也易从上式得到：

$$f_p(t) = f(t|\boldsymbol{\eta}) \sum_{m_i=1}^{\infty} m_i p_\theta(m_i) [S(t|\boldsymbol{\eta})]^{m_i - 1} \tag{3.2}$$

3.2 似然方法

下面推导统一治愈率模型的似然函数。设对于一个体 $i(i=1,2,\ldots,n)$，$Y_i=min(T_i,C_i)$ 是可观测的生存时间长度。其中 C_i 是数据发生右删失的时刻，具有随机性，并且与生存时间 T_i 相互独立。δ_i 是删失指示变量：当 $T_i \le C_i$ 时，$\delta_i=1$；当 $T_i > C_i$ 时，$\delta_i=0$。

再设 $\boldsymbol{x_i}=(x_{i1},x_{i2},\ldots,x_{ip})^T$ 是协变量向量。为了简化，将 n 维观测向量 $\boldsymbol{y}=(y_1,y_2,\ldots,y_n)^T$，$\boldsymbol{\delta}=(\delta_1,\delta_2,\ldots,\delta_n)^T$ 和 $\boldsymbol{m}=(m_1,m_2,\ldots,m_n)^T$ 与 $n\times p$ 维协变量矩阵 $X(x_1,x_2,\ldots,x_n)^T$ 一起定义：完整数据集 $D_c=(n,\boldsymbol{y},\boldsymbol{\delta},\boldsymbol{m},\boldsymbol{X})$，而无潜变量 $\boldsymbol{m}$ 的数据集 $D=(n,\boldsymbol{y},\boldsymbol{\delta},\boldsymbol{m},\boldsymbol{X})$。协变量 $\boldsymbol{X}$ 在模型中可用下式表示 $\theta\equiv\theta(\boldsymbol{x_i^T})\boldsymbol{\beta})$，其中 $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T$ 是回归系数向量。

如此，模型中的未知参数向量可由 $\boldsymbol{\phi}=(\boldsymbol{\beta^T},\boldsymbol{\eta^T})^{\boldsymbol{T}}$ 表示。可证完整数据集 D_c 的似然函数为：

$$L(\boldsymbol{\phi},D_c)=\prod_{i=1}^{n}[m_i f(y_i|\boldsymbol{\eta})]^{\delta_i}[S(y_i|\boldsymbol{\eta})]^{m_i-\delta_i}p_\theta(m_i) \tag{3.3}$$

上式含有不可观测的潜变量 $\boldsymbol{m_i}$，那么加总所有 $\boldsymbol{m_i}$ 可以得到观测数据的边缘似然函数，取对数后的表达式为：

$$l(\boldsymbol{\phi};D)=\sum_{i=1}^{n}\delta_i log[f_p(y_i|\boldsymbol{\phi})]+(1-\delta_i)log[S_p(y_i|\boldsymbol{\phi})] \tag{3.4}$$

3.3 PT 模型

当随机变量 M_i 服从参数为 θ 的 Poisson 分布时，统一治愈率模型就退化为提升时间模型（Promotion Time Model）[27]。PT 模型中参数 θ 与协变量的关系为：

$$\theta_i=exp(\boldsymbol{x_i^T}\beta) \tag{3.5}$$

因此随机变量 T_i 的生存函数和概率密度函数分别为：

$$S_p(t_i|\boldsymbol{\phi})=exp\{-e^{\boldsymbol{x_i^T}\beta}[1-S(t_i|\boldsymbol{\eta})]\} \tag{3.6}$$

和

$$f_p(t_i|\boldsymbol{\phi})=e^{\boldsymbol{x_i^T}\beta}f(t_i|\boldsymbol{\eta})exp\{-e^{\boldsymbol{x_i^T}\beta}F(t_i|\boldsymbol{\eta})\} \tag{3.7}$$

上式中第 i 个观测的治愈率由协变量表达：$p_{\theta_i}(0)=exp[-exp(\boldsymbol{x_i^T}\beta)]$。最后 PT 模型对数形式的边缘似然函数就是将前式的 $S_p(t_i|\boldsymbol{\eta})$ 和 $f_p(t_i|\boldsymbol{\eta})$ 替换成上式的表达形式。

3.4 服从 Weibull 分布的生存数据在 RACUF 累积和控制图的得分函数

对于 Weibull 提升时间模型，假定可疑个体的死亡时间服从前式的分布，并且协变量只对治愈率有影响，那么 $\gamma = 0$。

过程受控时，设随机变量 $R_{ij}(i = 1, 2, \ldots, n, j = 1, 2, \ldots, M_i)$ 服从 Weibull 分布，参数为 α 和 $\lambda = \lambda_0$。而对于失控过程，设参数 α 不变，只有 $\lambda_1 = \rho_1\lambda_0$ 变化。$0 < \rho_1 < 1$ 和 $\rho_1 > 1$ 分别表示 λ_1 变小和变大，也分别表示非免疫个体平均生存时间缩短和增加。

该模型的累积和控制图为:

$$W_i = l(\boldsymbol{\phi_1}; D_i) - l(\boldsymbol{\phi_0}; D_i) \tag{3.8}$$

其中 $\boldsymbol{\phi_l} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}_l^T)^T, l = 0, 1$ 是和 R_{ij} 的分布相关的参数向量，$\boldsymbol{\eta_0} = (\alpha, \lambda_0)$ 和 $\boldsymbol{\eta_1} = (\alpha, \lambda_1)$ 分别表示受控和失控状态下的检测参数值。这里的 D_i 表示第 i 个个体的信息。而将全部个体的信息加总可得到全数据的边缘似然函数，即：

$$l(\boldsymbol{\phi}; D) = \sum_{i=1}^{n} l(\boldsymbol{\phi}; D_i) \tag{3.9}$$

用 Weibull 分布拟合未免疫个体存活时间，生存函数和概率密度函数分别为：

$$\begin{aligned} S(t_i|\boldsymbol{\eta}) &= exp(-\frac{t_i}{\lambda}^{\alpha}), \\ f(t_i|\boldsymbol{\eta}) &= \frac{\alpha}{\lambda}(\frac{t_i}{\lambda}^{\alpha})^{\alpha-1} exp(-\frac{t_i}{\lambda}^{\alpha}) \end{aligned} \tag{3.10}$$

将上式带入前式，可得到长期生存函数 $S_p(t_i|\boldsymbol{\phi})$ 及其密度函数 $f_p(t_i|\boldsymbol{\phi})$。由此可以得到个体观测的对数似然函数：

$$l(\boldsymbol{\phi}; D_i) = \delta_i[x_i^T\beta - \alpha log(\lambda) + log(\alpha y_i^{\alpha-1}) - (y_i/\lambda)^{\alpha}] - e^{x_i^T\beta}[1 - exp(-(y_i/\lambda)^{\alpha}] \tag{3.11}$$

最后可得到 RACUF 累积和控制图的得分函数：

$$W_i = \delta_i[-\alpha log\rho_1 + (\frac{y_i}{\rho_0})^{\alpha}(1 - \rho_1^{-\alpha})] - e^{x_i^T\beta}\{exp[-(\frac{y_i}{\lambda_0})^{\alpha}] - exp[-(\frac{y_i}{\rho_1\lambda_0})^{\alpha}]\} \tag{3.12}$$

Weibull 模型参数的最大似然估计是最大化全数据似然函数 $l(\boldsymbol{\phi}; D) = \sum_{i=1}^{n} l(\boldsymbol{\phi}; D_i)$。而在实践中，我们利用 R 软件中 *optim* 实现。

考虑到计算有效性的问题，本文将部分参数进行了调整：17 式中的 $\alpha = exp\{\alpha^*\}$，$\xi = -\alpha log(\lambda)$。如此保证了所有的参数在估计过程中是有意义的。

第 4 章　模拟研究与实际应用

4.1　模拟研究

累积和控制图的表现情况常用平均运行时长（Average Run Length,ARL）衡量。所谓 ARL 是指控制图从检测开始到它发出生产出现问题的警报为止的抽取的平均样本组数，受控的 ARL 被称为 ARL_0，失控的 ARL 被称为 ARL_1，是用作检测过程是否发生实质变化，也可以作为检测速度快慢的参考指标。于此相关的概念是控制线，即当检测统计量落在控制线以外的区域时，控制图会给出过程失控的警报。检测一般的方法是将 ARL_0 固定，获得控制图的控制线 h，然后用相同的控制线 h 获得 ARL_1 衡量该控制图的效果。

为要更深入地研究 RAST 和 RACUF 累积和控制图，我们固定两表的 ARL_0，比较 ARL_1，越早发现变化的控制图表现越好。

下面使用带有和不带有治愈情况的两组样本进行模拟研究。对于带有治愈情况的样本，我们使用 RAST 方法研究监测速度和效果；而对于无法确定是否带有至于情况的样本，则使用 RACUF 方法进行研究。

4.1.1　思路

比较 RAST 和 RACUF 控制图的思路是比较两个控制图在检测过程发现失控，并且发出警报的样本的数目和速度。根据 Sego 等人的研究，设定一个受控的 ARL_0，累积和控制图的控制线 h 是可以通过模拟的方法近似获得的，那么使估计的 ARL_0 的 K 次重复平均值近似等于设定的 ARL_0，就可以获得控制线 h。

本次模拟中，随机产生一个容量为 100，服从一个含有协变量的分布，有删失情况，并且可能含有治愈情况的样本，利用这些样本作训练集，获得受控过程的参数 $\boldsymbol{\eta_0}$。

给定一个模型，上面参数的估计值被当作真实值使用，产生新的受控数据。其中 RAST 的模型使用的是 AFT 模型估计的参数值，RACUF 则使用 PT 模型估计的参数值。以此计算出两个控制图的值 $Z_i, i = 1, 2, \ldots, n$。对于两个控制图，记录下 $Z_i > h$ 的第一个位置，这就是过程发生警报的观测数目 ARL_1，更小的表明检测速度更快，而变化范围更小的则表明检测的比较准确。

本次模拟设 ARL_0=1000，根据两个模型的不同参数组合，得到估计的 ARL_0，从而找到最大控制线 h。具体来说，先产生 1000 个受控的样本，每个样本含有 10000 个观测，根据这些数据，最小化函数 $f(h) = |1000 - ARL_0|$ 得到 h 的估计值，其中 ARL_0 是那 1000 个样本中凡有警报 $(Z_i > h)$ 的平均观测数目。

为减轻计算负担，本文使用了黄金部分搜索法（Golden Section Search method）计算最大控制线 h[28]。采用这种方法，在模拟过程中未发出警报的样本将被去掉。由于尽可能多的样本更能保证最后结果的有效性，这样可能会出现偏差，但是模拟后的结果表明样本未发出警报的事件相当稀少，10000 个样本中只有约 12 个。因此采用这种方法是合理的。

最后检测本次模拟的检验效果，由每个含 5000 观测的 10000 个失控状态样本得到发

生警报的平均观测数目，即 ARL_1，并进行比较。

4.1.2 模拟环境

下面简述比较 RAST 和 RACUF 累积和控制图的模拟环境。首先由 Weibull 分布产生生存时间（T_i 和 R_{ij}），协变量是 $x_i(i = 1, 2, \ldots, n)$，服从参数为 0.5 的伯努利分布。对于没有治愈情况的数据，生存时间 T_i Weibull$(\alpha, \lambda e^{\gamma x_i})$。我们设定 $\alpha = 4$，$\gamma = -0.5$，对于受控数据设 $\lambda = \lambda_0 = 40$，而对于失控数据 $\lambda_1 = \rho_1 40$。对于含有治愈情况的数据，由均值 $\theta_i = e^{\beta x_i}$ 的 Poisson 分布生成潜变量 M_i，用以表示第 i 个个体可能有的风险因素数目，而不同的治愈率则是通过改变 β 的值实现的。对于第 i 个未治愈个体，随机样本 R_{ij} 样本容量为 M_i，由 $\alpha = 4$ 和 $\lambda = \lambda_0 = 40$ 的 Weibull 分布产生受控状态的样本，由 $\alpha = 4$ 和 $\lambda = \rho_1 40$ 产生失控状态的。这样死亡时间用 $t_i = minR_{ij}, j = 1, 2, \ldots, M_i(i = 1, \ldots, n)$ 表示。

用 6 个数表现 ρ_1 的敏感度，生存时间缩短的有 $\{0.5, 0.7, 0.9\}$，增加的有 $\{1.1, 1.3, 1.5\}$。删失时间 C_{ij} 服从 $[0, K]$ 上的均匀分布且彼此独立。为充分利用数据，取两个删失比例 50% 和 70%，这样就可以分析所有数据。

模拟研究中的删失比例可如下表示：$\frac{\text{样本中删失或治愈数目}}{\text{样本总数}}$

我们使用 AFT 模型拟合没有治愈情况的数据，治愈率为 50%($\beta = -0.77$) 由 PT 模型拟合。两种模型都要考虑删失比例为 50% 和 70% 两种情况。假定对于含治愈情况的数据，删失比例为 50% 时，所有删失个体都被治愈；删失比例为 70% 时，约 20% 个体可能死亡。观测时间和状态分别用 $y_i = min\{t_i, c_i\}$ 和 δ_i 表示，其中若 $t_i \le c_i$ 则 $\delta_i = 1$，否则 $\delta_i = 0(i = 1, 2, \ldots, n)$

我们使用 R 软件计算参数最大似然估计值的数值解，即用 R 内置函数 *optim* 计算 PT 模型的参数值，用 surbival 包中的 *survreg* 函数计算 AFT 模型的参数值 [29]。

4.1.3 结论

表 4.1 展示治愈率为 50%，RAST 和 RACUF 两个累积和控制图的模拟结果。两种情况未设置固定控制线，最大控制线 h 的变化依赖过程敏感度 ρ_1 的变化，数据的删失比例等因素。

由表 4.1 可知：RACUF 大多数的最大控制线 h 是 4 和 5，少数小于 3；而 RAST 控制线的变化更大，有一个大于 6，还有一个小于 2。再看 ARL 的表现：过程受控时，所有的 ARL 都接近设定值 1000，基本符合预期。由于 RACUF 的 ARL_1 比 RAST 的更小，变化也更小（特别是 $\rho = 0.9$ 时），所以 RACUF 效果更好，检测更敏感。

表 4.1　50% 治愈率：RAST 和 RACUF 模拟结果

删失比例	控制图类型	ρ_1	h	ARL_0	SE_0	ARL_1	SE_1
50%	RACUF	0.5	5.21	1021	3.13	8	0.01
		0.7	4.36	1032	3.42	14	0.03
		0.9	3.21	1023	3.65	76	0.14

续下页

续表 4.1　50% 治愈率：RAST 和 RACUF 模拟结果

删失比例	控制图类型	ρ_1	h	ARL_0	SE_0	ARL_1	SE_1
		1.1	2.76	1021	3.03	106	0.32
		1.3	3.53	994	3.13	13	0.02
		1.5	3.59	964	2.87	6	0.03
	RAST	0.5	3.35	1032	2.82	318	0.92
		0.7	1.94	998	2.81	561	1.42
		0.9	1.53	1032	2.21	826	1.78
		1.1	1.65	978	2.32	821	1.65
		1.3	2.48	1032	2.67	611	1.86
		1.5	4.36	1031	2.71	564	1.32
70%	RACUF	0.5	4.65	963	2.81	7	0.02
		0.7	4.35	1037	3.21	14	0.05
		0.9	3.61	943	2.87	73	0.16
		1.1	2.67	964	2.83	124	0.34
		1.3	3.75	982	3.04	37	0.06
		1.5	4.31	1034	3.27	26	0.08
	RAST	0.5	5.05	984	3.09	42	0.12
		0.7	4.25	1035	3.22	74	0.22
		0.9	3.98	1028	2.85	315	0.76
		1.1	2.46	956	2.65	365	0.86
		1.3	6.41	1025	3.18	145	0.45
		1.5	4.24	987	2.95	65	0.21

表 4.2 展示的是无治愈率数据的模拟结果，最大控制线 h 仍然受各种因素影响而变动。$\rho_1 = 0.9$ 时，大多数控制线接近 5，但这些值偏小。ARL_0 接近设定值 1000，也符合预期。但与含治愈情况不同的是，两图在检验生存时间变化时有详尽表现。

表 4.2　50% 治愈率：RAST 和 RACUF 模拟结果

删失比例	控制图类型	ρ_1	h	ARL_0	SE_0	ARL_1	SE_1
50%	RACUF	0.5	5.24	1021	3.41	5	0.01
		0.7	5.21	974	2.98	14	0.01
		0.9	3.61	1002	3.03	66	0.14
		1.1	1.94	1021	2.98	264	0.76
		1.3	4.32	1011	3.15	19	0.04
		1.5	4.59	964	3.07	14	0.03
	RAST	0.5	5.35	965	3.02	4	0.01

续下页

续表 4.2　50% 治愈率：RAST 和 RACUF 模拟结果

删失比例	控制图类型	ρ_1	h	ARL_0	SE_0	ARL_1	SE_1
		0.7	4.94	928	2.91	16	0.04
		0.9	3.53	962	2.91	82	0.18
		1.1	2.65	998	3.32	89	0.15
		1.3	3.48	992	3.07	21	0.06
		1.5	4.26	1011	3.71	14	0.03
70%	RACUF	0.5	4.65	963	2.81	7	0.02
		0.7	4.35	1037	3.21	14	0.05
		0.9	3.61	943	2.87	73	0.16
		1.1	2.67	964	2.83	124	0.34
		1.3	3.75	982	3.04	37	0.06
		1.5	4.31	1034	3.27	26	0.08
	RAST	0.5	5.05	984	3.09	6	0.12
		0.7	4.25	1035	3.22	14	0.02
		0.9	3.98	1028	2.85	85	0.16
		1.1	2.46	956	2.65	75	0.18
		1.3	6.41	1025	3.18	22	0.05
		1.5	4.24	987	2.95	15	0.03

由此可以得出结论：若数据含有治愈个体，那么在检测中 RAST 没有 RACUF 效率高，这是因为 RACUF 考虑了治愈个体的信息；而在检测不含治愈情况的样本时，RAST 和 RACUF 表现的基本一致，甚至在表 1 中差异最大的 $\rho_1 = 0.9$ 情况中，两种方法的表现也是基本一样的。

第 5 章　实际应用

下面引入实际数据研究 RACUF 累积和控制图的表现和与 RAST 累积和控制图的比较。该数据引自 Kersey 等人 1987 年利用骨髓移植方法治疗急性淋巴细胞白血病的研究。

5.1　数据说明与数据处理

对参与研究的两组患有白血病的 90 名病人进行分析，包括第一组的 46 名异源治疗病人和第二组的 44 名同源治疗病人。首先得到第一组病人的 Kaplan-Meier 图（图 5.1），根据图发现删失比例约为 20%，表明这些病人已经被治愈。再分别使用 RACUF 和 RAST 方法研究病人的生存时间的变化和两组病人生存时间的关系。根据常规做法，我们将第一组病人的部分数据作为训练数据，建立模型、估计参数并用以检测第一组的剩余病人和第二组病人的生存时间的情况 [30]。

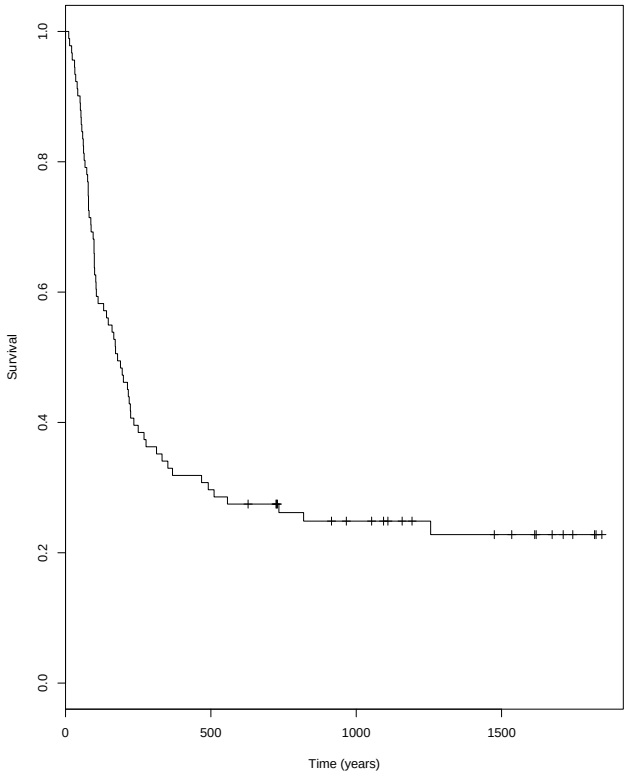

图 5.1　第一组病人的 K-M 图

按照本文的方法，我们用给定的协变量刻画每个个体的风险，分别用 RACUF 法的 Weibull 提升时间模型（3.12）和 RAST 法的 Weibull 加速死亡模型（2.6）拟合生存时间，调整受控与失控状态的参数比 ρ 来获得控制线 h。但实际计算量会很大，因此我们使用更加简单的重抽样方法，而不是通过 ARL_0 获取 h。根据 Oliveira 的模拟结果，重抽样是有

效的。

由于样本量较少，为保证两种方法的有效性，我们仍然采用模拟的方法，尽量使得模拟环境和真实数据的情况相近。模拟中的协变量由参数为 0.5 的伯努利分布随机生成，治愈率为 20%。用 RACUF 和 RAST 方法绘制出的累积和控制图见图 5.2 和图 5.3。

5.2 试验结果与分析

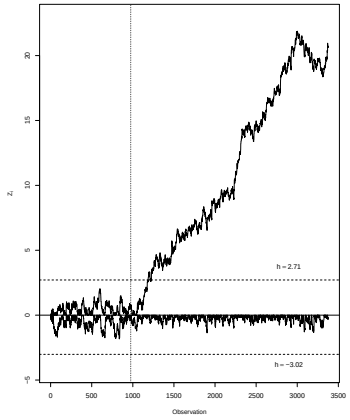

图 5.2 RACUF 累积和控制图

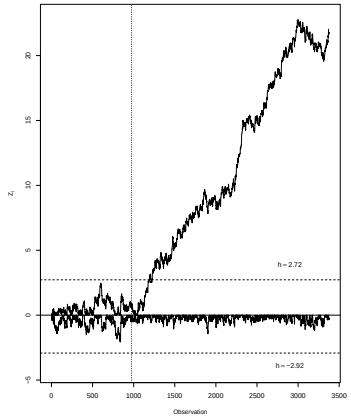

图 5.3 RAST 累积和控制图

两图中，0 以上的点表示平均生存时间增加，0 以下的点表示减少，即控制线 h 小于 0。水平虚线表示控制线，垂直点表明过程异常，发出警报的观测点。

两表均在病人的平均生存时间增加约 10% 时发出警报，总体上两图表现的差不多，但 RACUF 比 RAST 发出警报的时间要稍早一些，表明 RACUF 图更为灵敏。

第 6 章　总结与展望

6.1　论文工作总结

在研究中，部分病人被治愈的情况很少被研究者考虑在内，导致模型出现偏差，本文提供的含治愈情况的 RACUF 累积和控制图对原本的 RAST 累积和控制图进行了优化。

本文使用了 Weibull 提升时间模型给出了累积和控制图的得分函数，其中协变量只影响治愈率，而在模拟研究中比较 RACUF 和 RAST 两种方法，得出了检测含治愈情况的数据时，RACUF 比 RAST 表现更好；而在无治愈情况的情形中，两种方法表现类似。由于长期病患生存时间长，要求模型考虑治愈情况，并且要敏感于数据的变化，因此在不能确定数据是否含治愈情况时，使用 RACUF 累积和控制图更为合理。实际上只要治愈情况存在，无论治愈率多小，RACUF 都比 RAST 的检测速度快的多，准确度也更高。

另外两种方法的控制线 h 都随着数据和模型的变化而变化，因此在计算 h 时应当充分考虑到计算量过大的问题。本文的实际应用中，没有直接使用平均运行长度计算控制线 h，而是先通过重抽样的方法，如此在保证有效性的前提下简化了计算。

6.2　研究工作展望

本文的研究中仍有一些不完善之处，有待之后的继续研究：

1、尽管考虑到了患者被治愈的情况，但本文的方法只能检测受一个参数影响的患者，并且检测缺乏动态性，无法追溯患者的治愈率发生变化后的情况。实际上长期病人的生存时间往往受到多个参数的影响，比如治疗周期和治疗效果，如果在一个疗程中病人的反馈不佳，那么这些病人的治愈率就会减少。因此之后需要研究受更多参数影响的病人生存时间的检测情况。

2、本文并没有过多考虑估计误差问题，估计值被当作真实值使用。但这样并不合理，需要用不同的训练样本在假定模型中重复估计，这也是将来工作的重要课题。

3. 实际应用的原始数据过少，尽管尽可能采用和实际数据相同的环境进行模拟，但是仍然说服力仍显不足。如何获取更加有效的生存分析原始数据是下一步亟待解决的问题之一。

作者攻读学位期间发表的学术论文目录

[1] 郑为益, 傅红卓. 基于生存分析的客户流失模型研究 [D]. 华南理工大学,2011.

[2] Fan Jianqing and Lv Jinchi. A selective overview of variable selection in high dimensional feature space[J]. Statistica Sinica 2010,7(20): 101-148.

[3] Engler D. and Li Yi. Survival Analysis with High-Dimensional Covariates:An Application in Microarray Studies[J] Statistical Applications in Genetics and Molecular Biology. 2009,8(1):1-20.

[4] Variyath A.M., Chen Jiahua and Abraham Bovas. Empirical likelihood based variable selection[J]. Journal of Statistical Planning and Inference. 2010,5(140): 971-981.

[5] 彭非, 王伟. 生存分析 [M]. 北京: 中国人民大学出版社,2004.

[6] 王兆军, 邹长亮, 李忠华. 统计质量控制图理论与方法 [M]. 北京: 科学出版社,2013.8.

[7] Woodall.The Use of Control Charts in Health Care Monitoring and Public Health Surveillance[J].Journal of Quality Technology 38,2007.

[8] 王占锋, 应志良. 删失回归模型中若干统计问题的研究 [P]. 中国科学技术大学,2008.

[9] Leeb H. and Potscher B.M. . Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some generalower risk-bound results[J].EconometricTheory,2006,2(l): 69-97.

[10] 满敬銮, 许青松. 生存数据模型的变量选择 [D]. 中南大学,2009.

[11] Fan, J. and Li, R. Statistical Challenges with High Dimensionality:Feature Selection in Knowledge[C]. Proceedings of the International Congress of Mathematicians,2006.

[12] Wang H.S, Li R. and Tsai C.L. Tuning parameter selectors for the smoothly clipped absolute deviation method[J]. Biometrika. 2007,94(2), 553-556.

[13] Khoshgottaar T.M., Munson J.C. Predicting software development errors using software complexity metrics[J]. IEEE Journal on Selected Areas in Communications, 1990,8(2):253-261.

[14] Pickard L”Kitchenham B.and Linkman S. An Investigation of Analysis Techniques for Software Datasets[C]. In: Proceedings of the 6th International Symposium on Software Metrics, 1999: 130-142.

[15] Wang, Y. A New Approach for fitting Linear Models in high-dimensional Spaces[D].University ofWaikato, New Zealand, 2000.

[16] Giancarlo S., Milorad S.,Witold P. Advanced Statistical Models for Software Data[R],University of Alberta, 2003.

[17] Schneidewin N.F. Investigation of logistic regression as a discriminant of software quality[C].In: Proceedings of the 7th International Symposium on software metrics, 2001: 328-337.

[18] 董时富等. 生物统计学 [M]. 北京: 科学出版社,2002 年 8 月.

[19] Kersey, et al. Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia[J].New England Journal of Medicine,1987.

[20] Montgomery, D. C. Introduction to Statistical Quality Control[J].New York, NY,2007.

[21] Grigg, O. A., Farewell, V. T. and Spiegelhalter, D. J. Use of risk-adjusted CUSUM and RSPRTcharts for monitoring in medical contexts[J]. Statistical Methods in Medical Research 12, 147−170,2003.

[22] Jones, L. A., Champ, C. W. and Rigdon, S. E. The run length distribution of the CUSUM with estimated parameters[J]. Journal of Quality Technology 36, 95−108,2004.

[23] Moustakides, G. V. Optimal stopping times for detecting changes in distributions[J]. The Annals of Statistics 14, 1379−1387,1986.

[24] Sego, L. H., Reynolds, M. R. and Woodall, W. H. Risk-adjusted monitoring of survival times.[J] Statistics in Medicine 28, 1386−1401,2009.

[25] Rodrigues, J., Cancho, V. G., de Castro, M. and Louzada-Neto, F. On the unification of long-term survival models[J]. Statistics and Probability Letters 79, 753−759,2009.

[26] Chen, M. H., Ibrahim, J. G. and Sinha, D. A new Bayesian model for survival data with a surviving fraction[J]. Journal of the American Statistical Association 94, 909−919,1999.

[27] Yin, G. and Ibrahim, J. G. Cure rate models: a unified approach[J]. Canadian Journal of Statistics 33, 559−570,2005.

[28] Pan, X. and Jarrett, J. E. On quality control chart construction and simulation[J]. College of Business Working Paper Series, University of Rhode Island,2013.

[29] Therneau, T. and Lumley, T. Survival: survival analysis, including penalised likelihood. S original by Terry Therneau and ported by Thomas Lumley,2014. R package version 2.37-7. R Foundation for Statistical Computing, Vienna. CRAN. Available at: <http://cran.r-project.org/web/packages/survival/index.html>.

[30] Maller, R. A. and Zhou, X. Survival Analysis with Long-term Survivors[J]. Wiley, New York, NY,1996.

致　谢

时光荏苒，不知不觉间我的研究生生涯已近尾声。面对刚刚完成的硕士论文，我的心中充满了喜悦。值此论文完成之际，我要感谢所有在我攻读硕士学位期间炮经无私帮助过我的人！

衷心感谢我的导师胡涛副教授！三年求学生涯中，导师的科研精神使我获益匪浅，受用终生！这篇硕士论文是在导师的精心指导和鼓勘下才完成的，从论文选题、查阅文献、收集资料直至最后的论文撰写和审阅都倾注了导师的智慧和心血。祝导师身体健康，万事如意，合家幸福！

衷心感谢各位在我研究期间给与过我帮助的各位同学，在你们的帮助下，我才得以成功的完成这篇硕士论文，在此向你们表示感谢，祝福你们前程似锦！

衷心感谢我的父母和其他家人，是你们无私的爱，让我懂得了什么是责任和感恩！

最后衷心感谢在百忙之中参加评阅论文的各位专家和教授！