A Comparative Inference Based on the Median Survival Time in Two-Sample Problem

Taro Amagasaki

MARCH 2010

A Comparative Inference Based on the Median Survival Time in Two-Sample Problem

A dissertation submitted to
THE GRADUATE SCHOOL OF ENGINEERING SCIENCE
OSAKA UNIVERSITY
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN ENGINEERING

BY

Taro Amagasaki

MARCH 2010

謝辞

本論文の作成におきましては，多くの方々にご指導・ご支援を頂戴いたしました．ここに，深くお礼を申しあげます．

指導教官の白旗慎吾先生には，本論文の全体を通して，多大なご教示をいただきました．来阪した際には，いつも優しくご指導していただきました．また，様々な会合でお会いしたときには，笑いを誘う和やかな雰囲気のなか，いつも温かいお話を聞かせていただきました．心よりお礼を申し上げるとともに，今後ともますますのご高配のほどよろしくお願い申し上げます．大阪大学教授（大学院基礎工学研究科　統計数理講座）の狩野 裕先生には，公聴会にて大変貴重なご指摘を賜りました．ご指摘のなかには，将来の研究課題にもつながる貴重なご意見もいただきました．大阪大学教授（大学院基礎工学研究科　統計数理ファイナンス講座）の内田 雅之先生には，公聴会の冒頭にて大変緊張していた私を気に留めて，緊張をほぐすようなコメントをいただきました．また，大変率直なご指摘をいくつも賜り，筆者の研究内容のみならず，発表のスキルについても考えさせる機会を与えていただきました．大阪大学准教授の坂本　亘先生には，本博士論文の準備に際して多大なご配慮をいただきました．心よりお礼を申し上げます．

本博士論文の作成にあたり，医学統計研究会（Biostatistical Research Association，通称 BRA）の皆様には多大なご支援を頂戴いたしました．とくに，大阪大学医学部助教の杉本 知之先生には，これ以上ないほどのご支援を賜りました．研究の内容や研究に対する姿勢のみでなく社会人としての姿勢にも大変厳しくも温かいご支援を賜りました．この博士論文は杉本先生のご支援なくして出来上がらなかったものであると痛感している次第でございます．この場を借りて，心よりお礼を申し上げるとともに，今後ともご指導ご鞭撻のほどよろしくお願い申し上げます．

後藤 文彦さんには，お酒の席をご一緒させていただいた折に，様々なことを学ばせていただきました．とくに，人とのつながりを大切にし，また故郷，とりわけ両親への感謝を忘れてはならないと教えていただいたことが印象に残っております．イーピーエス（株）の魚井 徹博士には，学会などでお会いした際には，製薬企業で働いている立場でのご意見などをたくさん頂戴いたしました．ご臨床研究情報センター [財団法人　先端医療振興財団] の松原 義弘博士には，私がシオノギ製薬（株）に勤務していた頃はもとより，学生の時分から多大なご支援を賜りました．いつも温かく私を見守っていただいたことはもと

より，いつも私の体を気遣っていただきました．いつも煙草を片手に様々なお話を出来たことは今の私の糧となっております．今後とも，ぜひ遊学をご一緒させていただければと思っております．フィールドワークス（株）の木田 義之さんには，諸種の会合で幾度も激励のお言葉をおかけいただき，また“ 遊 ”にも幾度も誘っていただきました．ソリューションラボ（株）の志賀 功さんには，筆者が大分統計談話会に参加した際にはいつも大変にお世話になりました．富士通ソフトウェア・ラボラトリ（株）の衛藤 俊寿博士には，大分統計談話会の折にはいつもお世話になり，筆者に大分の良さをたくさん教えていただきました．ファイザー（株）の栗林 和彦博士には，BRA の会合などでお会いした際に筆者の研究内容に関わらず大変貴重なご意見を頂戴いたしました．第一三共（株）の佐藤 俊之博士には，ご自身の博士取得時のご経験などを熱く語っていただきました．また，アメリカへの長期出張など，その仕事への積極的な姿勢にも大変啓発されました．小野薬品工業（株）の冨金原 悟博士には，遊学ともに，大変お世話になりました．ファイザー（株）の河合 統介博士には，様々な場面で遊学をご一緒し，いつも筆者をかわいがっていただきました．河合さんの研究に対するダイレクトな姿勢には尊敬するものがあり，筆者の努力のなさを何度も痛感させられました．また様々な“ 遊び ”に精通されており，その遊びに対する姿勢を今後は学ばせていただきたいです．あすか製薬（株）の藤澤 正樹博士には，遊学ともに様々な場面でご一緒させていただきました．ご自身が，BRA および大愚の会の運営にかける姿勢には常に称賛と尊敬の念を抱いておりました．エーザイ（株）の高瀬 貴男さんには，常日頃から筆者を応援していただきました．とくに，高瀬さんには夜のお酒の席を何度もつくっていただき，筆者の日頃の疲れを忘れさせていただきました．協和発酵キリン（株）の古川 泰伸さんには，後藤研修士課程の同期として公私にわたり大変お世話になりました．また，博士課程でもともに切磋琢磨することができ，おおいに刺激を受けました．アステラス製薬（株）の伊藤 雅憲博士には，筆者の博士論文作成の準備に際して大変貴重なコメントを幾度となく頂戴いたしました．また，遊学のバランスに優れているその姿勢には学ぶことが多く，筆者の見本でした．ノバルティス・ファーマ（株）の池田 公俊さんには，同じ会社で働いてることもあり，研究・仕事に関わらずたくさんのお話を聴かせていただきました．アスビオファーマ（株）の永久保 太士さんには，同じ博士課程の同期として，たくさんの刺激を互いに共有できました．また，たくさんの楽しいお酒の場を提供してくださいました．これらの方々に重ねてお礼を申し上げます．

長崎大学教授の柴田 義貞先生には，BRA の諸会合などで直接お話する機会を通して，大変に啓発されました．柴田先生の科学者としての姿を幾度となく見る機会があり，大変

な感銘を受けました．大分大学教授の越智 義道先生には，BRA の諸会合などで大変貴重なコメントを頂戴いたしました．教育の現場でお忙しいにもかかわらず，セミナーなどでご自身の発表準備をされるお姿には大変学ぶべきことがたくさんありました．鹿児島高等専門学校教授の藤崎 恒晏先生には，いつも明るくまた温かく接してくださいました．藤崎先生のいつも元気ではつらつとした姿には大変元気づけられました．兵庫医科大学講師の大門 貴志先生には，ベイズ流接近法などについて様々なご支援を賜りました．また，本博士論文の執筆中には温かい励ましのお言葉を頂戴いたしました．山梨大学助教の下川俊雄先生には，様々な会合でお会いした際には，統計科学に関する話題のみならず様々なお話を聴かせていただき，大変元気づけられました．

ノバルティス・ファーマ（株） オンコロジー開発統括部 統計解析グループのグループマネージャーの谷口 淳介さんには，大阪大学 大学院基礎工学研究科 博士後期課程（社会人コース）に入学する機会を与えていただきました．この 3 年間研究を進めるにあたり，様々な困難にぶつかりましたが，その都度，筆者へのご配慮をいただきました．特に，業務が非常に忙しいときには，幾度となく温かいお声をかけていただき，また体調などについても気遣っていただきました．さらに，業務と研究をバランスよく行うのに，たくさんのアドバイスも頂戴いたしました．谷口さんのこれまでのご配慮に心よりお礼を申し上げます．

医学統計研究会の理事長の後藤 昌司先生には，筆者が大学院修士課程に在籍している際に本研究テーマを与えていただきました．後藤先生のもとで，生存時間解析の基礎から学び，また本研究テーマがどれほど製薬会社の実地において重要であるかを教えていただきました．また後藤先生と杉本先生の最強タッグのもと，たくさんの叱咤激励を頂きながら，本博士論文を仕上げるまでに至ったことに大変感謝しております．さらに，修士課程の時分からの教えである「掃除・勤行・学問」がいかに大切であるか身に染みた 3 年間でありました．まだ社会人としても科学者としても未熟である私に多大なご指導を賜りました．後藤先生から頂いた言葉で特に印象深いものは「全ての根源は“ 時間 ”である」と「“ 変人 ”にはなるな」でした．これらの言葉は筆者が人間としての魅力を醸成するための教えとして肝に銘じておく所存でございます．後藤先生とお会いしてからもう 10 年以上経ちますが，統計科学者としてまた人としての良い師匠に出会えたと深く感じております．ここに，今までのご指導に心より感謝いたします．大変ありがとうございました．

筆者は，博士後期課程（社会人コース）の 3 年間にわたり，ノバルティス・ファーマ（株）から経済的な支援を受けました．ここにその支援に対し深く感謝いたします．

最後に，筆者の身を絶えず心配し，励ましてくれた松江の祖父と両親に感謝いたします.

Abstract

Human beings sometimes experience a variety of events in the life, such as start of smoking, pregnancy and childbearing in female, and death in conformity with nature, as examples. We are often interested in the time to such an event of interest, and are able to handle time to event data in the framework of survival analysis. One of the main objectives of survival analysis is to compare two or more survival distributions. Here, we consider using of median survival time as the criterion of the comparison. The reasons why we prefer to evaluate the median are that some advantages such as cut of sample size and cost, and shortening of duration of clinical trial by achieving the conclusion early for the comparison, are expected. For the comparison of survival distributions, existing rank tests such as log-rank test have been frequently used in several areas. However, to maximize the use of rank test, we usually require the long follow-up to observe non-censored survival data as much as possible. Meanwhile, median can be estimated accurately compared to other statistics like mean survival time even if censored observations are involved, and it does not require the long follow-up, that is, we just need to observe survival data until the estimable time-point of median. As a major median test for right censored data, Brookmeyer and Crowley (1982b) extended the sign test procedure to the version of censored data. This median test is asymptotically valid. However, that median test does not consider the survival information after the median survival time, so it may cause the low power, especially in small sample. To overcome such problem, we proposed alternative median test procedure based on the property of order statistics in the framework of two-sample problem. In this thesis, we discussed the two-sample difference between the mid order statistics in order to conduct the median test based on estimating the significance probability. Furthermore, we provided the rationale to estimate the significance probability, a manner to cope with censored observations and some contrivances to overcome computational problem in that estimation. As a result, the null distribution of the proposed median test was asymptotically valid, and was investigated to be appropriate with a conservative tendency in the finite sample by the simulation. Also, simulation studies showed that the proposed test has the same or higher power than the existing median test procedures. Finally, we discussed the usefulness of the proposed test via case studies.

Notations

Notations	Definitions and examples	Remarks
General		
O_p	$O_p(n)$	Order of random variable
$B(\cdot,\cdot)$	$B(m,n)=\int_0^1 x^{m-1}(1-x)^{n-1}dx$	Beta function $(m,n>0)$
T_{ji}	$T_{ji}=\min(X_{ji},C_{ji})$	Observed survival time
X_{ji}		Non-negative continuous random variable
C		Censored time
S_j		Survival function of sample j
f_j		Probability density function of sample j
M_j		True median survival time of sample j
H_0		Null hypothesis: $M_1=M_2$ and $S_1(t)=S_2(t)\ \forall\ t$
$\mathrm{H}_0^{\mathrm{m}}$		Null hypothesis: $M_1=M_2$
M_0		Common median under H_0
M_0^*		Common median under $\mathrm{H}_0^{\mathrm{m}}$
n_j		Sample size for sample j
n	$n=n_1+n_2$	Pooled sample size
Generalized Sign Test		
T_{HS}		Sign test statistic for complete data
$\widehat{S}_j$		Kaplan-Meier estimate for sample j
$\widehat{S}_0$	$\widehat{S}_0(t)=n^{-1}\{n_1\widehat{S}_1(t)+n_2\widehat{S}_2(t)\}$	Weighted Kaplan-Meier estimate
$\widehat{S}_j^{\mathrm{lin}}$		Estimated survival probability found by linear interpolation
$\mathcal{N}_j(t)$		Counting process
$\mathcal{Y}_j(t)$		At risk process
T_{BC}		Test statistic
$T_{\mathrm{BC}}^{'}$		Modified test statistic
Empirical Likelihood Ratio Test		
$\widehat{\lambda}_j$	$\widehat{\lambda}_j(t)=d\mathcal{N}_j(t)/\mathcal{Y}_j(t)$	Unconstrained hazard
$\widehat{\lambda}_j^*$		Constrained hazard under $\mathrm{H}_0^{\mathrm{m}}$
$\log L_{\mathrm{u}}$	$\log L_{\mathrm{u}}(\widehat{\lambda}_j)$	Unconstrained maximum log-likelihood
$\log L_{\mathrm{c}}$	$\log L_{\mathrm{c}}(\widehat{\lambda}_j^*)$	Constrained maximum log-likelihood
T_{ELR}		Test statistic
$\widehat{S}_j^{\mathrm{c}}$		Constrained Kaplan-Meier estimate
α_j		Lagrangian parameter

Notations (continued)

Notations	Definitions and examples	Remarks
Bootstrap Median Test		
$\widehat{p}_{boot}$		Bootstrap p-value
Proposed Median Test		
Y		Random variable of difference in two-sample medians
$g_1(y)$		Density function of Y in complete data
$pv_1(\cdot)$		Significance probability (p-value) function
$\tilde{S}_j$		Discrete approximation of S_j
$\widetilde{pv}_1(\cdot)$		Discrete approximation of $pv_1(\cdot)$
$\widehat{g}_1(y)$		Density function of Y in censored data
$\widehat{pv}_1(\cdot)$		$\widetilde{pv}_1(\cdot)$ for which $\tilde{S}_j$ is replaced by $\widehat{S}_0$
γ_{21}	$\gamma_{21} = n_2/n_1$	Ratio of sample size between sample 1 and 2

Abbreviations

NNT	Number Needed to Treat
B&C	Brookmeyer and Crowley
ELR	Empirical Likelihood Ratio (Test)
p.d.f	Probability Density Function

Acknowledgment

The author would like to acknowledge the advices and kind helps of many people to the completion of this thesis. First of all, the author wishes to thank Professor Shingo Shirahata of Osaka University, who has provided adequate and useful suggestions throughout writing this thesis. The author is also deeply grateful to Professor Yutaka Kano and Masayuki Uchida of Osaka University for providing valuable suggestions. And, the author wishes to thank Dr Tomoyuki Sugimoto of Osaka University who has provided many helpful supports to me. Especially, thanks to his continuous helps, we have been able to continue the work since my college. I would like to take this opportunity to thank Dr Tomoyuki Sugimoto for his considerable helps since 1999.

The author wishes to thank the following members of the Biostatistical Research Association (BRA) because they have provided valuable motivation, feedback and ideas: Mr. Fumihiko Goto, Dr. Tohru Uwoi, Dr. Yoshihiro Matsubara, Dr. Norisuke Kawai, Dr. Masaki Fujisawa and Dr. Takashi Daimon. The author shows his special gratitude to Professor Yoshisada Shibata of Nagasaki University and Professor Yoshimichi Ochi of Oita University for giving valuable comments and suggestions. The author also expresses his gratitude to Mr. Takao Takase and Mr. Yasunobu Furukawa for supporting me as a friend.

The author has been able to devote so much time to his thesis because of the cooperation of his group manager Mr. Junsuke Taniguchi and my colleagues, in Novartis Pharma Oncology. Especially, Mr. Junsuke Taniguchi has provided many kind suggestions to find balance in author's life.

Dr. Masashi Goto, who is the representative of the Biostatistical Research Association (BRA), has provided precious advice not only on the research but on knowledge for living as a member of society. The author would especially like to thank him for giving valuable ideas and motivations for this work.

Finally the author is grateful to my parents for warm-hearted encouragement and kind support.

Contents

1. Introduction

1.1 Background

Human beings sometimes experience a variety of events in the life, for example, become a particular illness, start smoking, pregnancy and childbearing in female, and death in conformity with nature. In survival analysis, the time till an experience of particular event from a common starting time is referred to as a response. And, survival analysis makes it possible to summarize the survival information based on the particular event for one or more groups. Typically the elapsed time till an event of interest and the event itself are considered the survival time and death (or failure) in a comprehensive meaning, respectively.

One of the important objectives in survival analysis is to compare two or more groups, that is, to compare two or more survival distributions. Here, our concern is to evaluate the statistical difference in reduced survival experiences based on the individual survival time brought by several factors. An important problem in practice is "What measure should be evaluated for the difference in survival experiences, namely, survival distributions?".

For the comparison of survival distributions, consderable statistics as evaluation criterion are NNT (Number needed to treat), difference in survival rates (or mortality rate), five-year survival rate, median survival time and mean survival time. The difference in survival rates is a statistics which suggests the difference in survival probabilities at given a time point, therefore it provides different result depending on the time point of interest. Five-year survival rate is also a statistics providing only survival information at a specific time point, namely, five years. This statistics is used frequently in the area of cancer therapy, however, it does not make sense to evaluate 5-year survival for all cancer types. Thus, we need to take care of handling such statistics for the comparison of survival

distributions since we must lose the survival information other than one at single time point. NNT is defined as the reciprocal of the difference in survival rates (or mortality rates) at a time point. NNT has also the similar characteristics from a point view of the point estimation. Since these statistics are based on the point estimations of survival rates or mortality rates, the comparison of survival distributions based on these statistics is equivalent to the comparison of survival information at a time point. That is, these statistics does not reflect the overall survival information. In contrast, median survival time and mean survival time are considered to be useful compared to those statistics. Each statistics has the following characteristics in practice:

Mean survival time

- Advantages:
  - For the comparison of survival distributions, a group with the largest sum of survival times can be considered to have highest survival possibility. That is, mean survival time is proportional to the sum of survival times. In other words, mean survival time includes all information on the length of obtained survival times for a group.
  - It is easy to understand mean survival time intuitively for non-statisticians, namely, medical doctors and patients in medical practice.
- Disadvantages:
  - Mean survival time relies heavily on the shape of survival distribution since it is defined as the area under the survival distribution. Extremely speaking, even if survival distributions have different shapes, the mean survival time could be the same.
  - If the largest observed survival time is censored, then Kaplan-Meier curve does not reach at time-axis. Under this, it is very difficult to estimate mean survival time since we can't know the true shape of right tail of survival distribution exactly.

Median survival time

- Advantages:

  - As survival distribution is skewed with long tail on the right side in general, mean survival time can be larger than median survival time due to small number of long survivors. Thus, median survival time is robust statistics in that sense.
  - Median survival time is not subject to the influence of censored observations compared to the mean survival time.

- Disadvantages:
  - For the comparison of survival distributions based on median survival time, we can perform median test only when median survival time in all samples can be estimated. That is, if median in at least one sample can not be estimated, then we are not able to apply any of testing procedures based on the median survival time.
  - Median survival time includes the information of order mainly.
  - In clinical trial, median survival time is not sensitive to larger (or smaller) survival times even if they reflect significant benefit (or risk) of drug.

It is difficult to determine the best statistics for the comparison of survival distributions since each has advantages and disadvantages, respectively. The preferable statistics to be evaluated must be chosen depending on the applied case. However, we would like to describe several reasons why we focus on the median survival time in this thesis. First, we state its availability caused by the reduction of the overall survival distribution in clinical trials, where the "reduction" implies to be representative of all the information obtained from the survival distribution. The reduction of the survival distribution by median survival time, if possible, can shorten the duration of clinical trial, and thus lead to cut the cost and sample size. Next, it is very easy to interpret the median survival time as the 50 percentage assurance timepoint for the occurrence of event such as death compared to other percentiles. From a viewpoint of these possible contributions caused by evaluating the median survival time, we prefer to evaluate median survival time under a situation where two survival distributions are compared based on it.

When we consider comparing two or more samples, two-sample problem or multiple-sample problem will be open for discussion. Unlike two-sample problem, the objective

of comparison between samples in multiple-sample problem become diversified because order relation between samples and existence of a standard sample for comparisons should be considered. Tsubaki and Fujita (1987) suggested that the majority of clinical trials conducted in Japan are based on two-sample comparison even if number of samples of interest in a clinical trial are greater than or equal to 3. In summary, even in clinical trials having primary objective to compare ≥ 3 samples in parallel, two-sample test for each of considerable combinations of samples has been repeated since cut of required sample size and reduction of efforts to set up a number of clinical trials have been usually desired due to limited cost. That is, their suggestion is that multiple-sample comparison can be reduced to two-sample comparisons. In contrast, two-sample problem can be extended to multiple-sample problem in such sense. Because of this, we would like to focus on two-sample comparison only in this thesis.

1.2 Objectives of our works

Due to characteristics of skewed survival distribution and being censored observations, nonparametric tests such as the rank test have been frequently used than tests based on the difference in means (for example, t-test) for the comparison of survival distributions. One of the major advantages of the rank test is that they can reflect the entire survival experiences from survival distributions. On the other hand, they have several drawbacks such as requirement of long follow-up period to collect complete survival information. That is, it is desirable to collect non-censored survival information as possible in order to estimate survival distribution more exactly. To avoid long follow-up period for the purpose of application of rank tests, we consider that inference based on the difference in median survival times which could be robust statistics for various distribution forms would be very useful as a modified version of the method based on the means for the skewed survival distributions. We also know that an evaluation for the extension of survival potential based on the median survival time is not always powerful for any alternative hypothesis, while it is more powerful than rank test procedure under the alternative hypothesis with shift of the location (Brookmeyer and Crowley, 1982b).

There are relatively few papers which propose the testing procedure to compare two

or more samples based on the median survival time while there are many papers on the point or interval estimations of the median survival time in single sample with right-censored data. Bartholomew (1957) provided an interval estimate for the median survival time under exponential survival distribution. Nonparametric methods for constructing confidence interval have been proposed by Brookmeyer and Crowley (1982a), Emerson (1982), Simon and Lee (1982), Slud *et al.* (1982) and, Wang and Hettemansperger (1990). Slud *et al.* (1982) provided the "reflected" confidence interval, and they distinguished the method from "test-based" confidence interval which was proposed by other authors. They also went into the details of the difference between those two confidence intervals. Efron (1981) and Reid (1981) proposed an inference for the median survival time based on the bootstrap method.

Brookmeyer and Crowley (1982b) provided k-sample median test as an extension of the sign test by Hájek and Sidák (1967) to censored data version. Their median test is referred as the generalized sign test for right-censored data. Following the existing median test for complete data, they proposed to use the weighted Kaplan-Meier estimate in order to define the pooled-sample median survival time under null hypothesis. Then, their median test compares the Kaplan-Meier estimate for either one of two samples at the pooled-sample median survival time and survival probability of 0.5 under null hypothesis. This generalized sign test is a member of the class of Mood type median test while Gastwirth and Wang (1988) developed a median test included in the class of control median test. Both classes of median test are the same in that the difference between the estimated survival probability at the median survival time under null hypothesis and survival probability of 0.5 is evaluated. However, we note that the estimation method for the median survival time under null hypothesis is different. Test statistics in both classes do not have exchange invariance for two samples, that is, both classes of median test do not provide the same test statistic value when the standard distribution to be evaluated in the test is replaced to another one. Any modification to hold the exchange invariance has never seen in existing papers, but we present a proposal for the modification later briefly. These classes of median tests by Brookmeyer and Crowley (1982b), and Gastwirth and Wang (1988) seem to be asymptotically valid, but several issues such as expected low power due to the loss of survival information post median survival time, and the validity

of null distribution under small sample size should be considered.

Naik-Nimbalkar and Rajarshi (1997) proposed the empirical likelihood ratio test for the equality of k medians in right-censored data. As an important characteristic, their empirical likelihood approach does not consider a null hypothesis of equal survival distributions. Their empirical likelihood ratio test is based on a null hypothesis of equal median survival times. They considered that the empirical likelihood approach is the natural way for the handling of censored data less than the pooled-sample median survival time and the definition of the pooled-sample median survival time those which have been issued by Brookmeyer and Crowley (1982b). They showed that the empirical likelihood ratio statistic has chi-squared distribution with $k-1$ degrees of freedom under null hypothesis of equal medians. Several empirical likelihood approaches have been developed in the survival analysis so far. And the approach has many preferable properties such as its ability to carry out a hypothesis testing and construct confidence intervals without estimation of the variance. However, suffice it to say to that the test statistic is constructed based on the property of approximation. In Naik-Nimbalkar and Rajarshi (1997), they did not provide any numerical investigation for proposed empirical likelihood ratio test.

Park and Na (2000) proposed a bootstrap median test for right-censored data under two-sample problem. In their paper, the difference between the control median test and the mood type median test is clarified, and the Behrens-Fisher problem is also discussed briefly. To apply the bootstrap method to the testing procedure, they proposed to use the difference of medians as the control median test statistic directly, and they also provided the validity to use the asymptotic bootstrap distribution of it theoretically. The bootstrap median test consists of the bootstrap sampling from the combined sample to estimate the difference of medians under two bootstrap samples; estimating a p-value by counting the number of detections for which the difference of medians based on the bootstrap samples is greater than an actual observed difference of medians. The bootstrap median test can be an alternative procedure based on the approximate bootstrap distributions against the tests which assume the asymptotic normality based on the large sample approximation. Although the bootstrap median test is very simple to be applied, it requires a heavy workload on the computation of p-value.

Amagasaki *et al.* (2009) proposed a two-sample median test for right-censored data

based on the property of order statistics. They defined the density for the difference in two-sample mid order statistics, and then proposed to obtain the p-value through the significance probability (p-value) function directly. However, the integral calculation seems to be rather difficult even if any parametric survival distribution is assumed. Therefore they provided some ideas to overcome such difficulties in computing the p-value. Indeed, their median test has several difficulties in computing the p-value, but it is very natural manner to evaluate the difference of observed medians exactly.

As far as we know, there is no papers which describe the difference of performances for those median tests with right-censored data. Therefore, we carry out simulation and case studies in order to discuss how these median tests are different from their performances.

1.3 Outline of datasets

In this section, we present two datasets to be used for case studies in the later section.

[Data set No.1: Time to death for patients with cancer of the tongue($N = 80$): Sickle-Santanello *et al*, 1988]

A study was conducted to investigate the effects of ploidy on the prognosis of patients with cancers of the tongue (mouth). Patients were selected who had a paraffin-embedded sample of the cancerous tissue taken at the time of surgery. Follow-up survival in week data was obtained for each patient. The tissue samples were examined using a flow cytometer to determine if the tumor had an aneuploid (abnormal) or diploid (normal) DNA profile using a technique discussed in Sickle-Santanello *et al*. (1988).

Sample 1 consisted of 52 patients with aneuploid tumors, and sample 2 consisted of 28 patients with diploid tumors. Each had 21 and 6 censored observations, respectively.

[Data set No.2: Time to death for patients with gastric cancer($N = 90$): Stablein and Koutrouvelis, 1985]

A clinical trial of chemotherapy against chemotherapy combined with radiotherapy in the treatment of locally unresectable gastric cancer was conducted by the Gastrointestinal Tumor Study Group (1982). In this trial, 45 patients were randomized to each of the two arms and followed for about eight years. Survival data in days was reported in Stablein

and Koutrouvelis (1985).

Sample 1 consisted of 45 chemotherapy only patients, and sample 2 consisted of 45 chemotherapy plus radiotherapy patients. Each had 2 and 6 censored observations, respectively.

Table 1.1. Survival data for patients with tongue cancer

Aneuploid Tumors												
1	3	3	4	10	13	13	16	16	24	26	27	28
30	30	32	41	51	61+	65	67	70	72	73	74+	77
80+	81+	87+	87+	88+	89+	91	93	93+	96	97+	100	101+
104	104+	108+	109+	120+	131+	150+	157	167	231+	240+	400+	
Diploid Tumors												
1	3	4	5	5	8	8+	12	13	18	23	26	27
30	42	56	62	67+	69	76+	104	104	104+	112	129	176+
181	231+											

+ censored observations

Table 1.2. Survival data for patients with gastric cancer

Chemotherapy only									
1	63	105	129	182	216	250	262	301	301
342	354	356	358	380	383	383	338	394	408
460	489	499	523	524	535	562	569	569	676
748	778	786	797	955	968	1000	1245	1271	1420
1420	1694	2363	2754^+	2950^+					
Chemotherapy plus Radiotherapy									
17	42	44	48	60	72	74	95	103	108
122	144	167	170	183	185	193	195	197	208
234	235	254	307	315	401	445	464	484	528
542	547	577	580	795	855	1366	1577	2060	2412^+
2486^+	2796^+	2802^+	2934^+	2988^+					

+ censored observations

2. Competitive Median Tests for Right Censored Data

In this chapter, we introduce some existing median tests for right censored data. Before introducing existing median tests, we discuss nonparametric Behrens-Fisher problem first. By discussing the problem, a problem for selecting a null hypothesis to be tested is issued. Then, we introduce three existing median tests for right censored data, namely, the generalized sign test, empirical likelihood ratio test and bootstrap median test. For the empirical likelihood ratio test, no numerical investigation was provided in the paper (Naik-Nimbalkar and Rajarshi, 1997). Therefore, we propose a manner how to apply the empirical likelihood ratio test to survival data practically. Finally, we propose a median test based on the property of order statistics. Here, we discuss the two-sample difference between mid order statistics, so that we conduct the median test based on the estimation of the significance probability function. For this test procedure, we provide the rationale to estimate the significance probability, a manner to cope with censored data and a few contrivances to overcome computational problem in such estimation.

2.1 Null hypotheses and Behrens-Fisher problem

Let $S_j(t)$ and M_j^* be the true survival function and the true median in the sample j, respectively $(j = 1, 2)$. One of the considerable null hypotheses is

$$\mathrm{H}_0^{\mathrm{m}} : M_1^* = M_2^*,$$

that is, two medians are equal. However, we can not always assure $S_1 = S_2$ even if $\mathrm{H}_0^{\mathrm{m}}$ is true. Thus, it is referred as nonparametric Behrens-Fisher problem. If $\mathrm{H}_0^{\mathrm{m}}$ is rejected, we can conclude $S_1(t) \neq S_2(t)$ at t in the neighborhood of median. A candidate of null

hypothesis to address the nonparametric Behrens-Fisher problem is

$$\mathrm{H}_0 : M_1^* = M_2^* \text{ and } S_1(t) = S_2(t) \ \forall \ t.$$

Thus one can avoid a null hypothesis $\{M_1^* = M_2^* \text{ and } S_1 \neq S_2\}$ as a part of $\mathrm{H}_0^{\mathrm{m}}$ by adopting H_0. For simplicity, H_0 can also be defined as $\mathrm{H}_0 : S_1(t) = S_2(t)$ $(\forall\ t)$ since equal survival distributions must have a common median. If one is interested in the Neyman-Peason hypothesis testing procedure, one shall set an alternative hypothesis $\mathrm{H}_1 : S_1(t) \neq S_2(t)$ against H_0. However, we need to note that $\mathrm{H}_1 : S_1(t) \neq S_2(t)$ consists of $\{M_1^* = M_2^* \text{ and } S_1 \neq S_2\}$ and $\{M_1^* \neq M_2^* \text{ and } S_1 \neq S_2\}$.

Even if our interest is to test equality of two medians, we note that existing median tests for right-censored data have not been developed with consistent null hypothesis. Therefore, we need to consider used null hypothesis carefully when we perform median test.

2.2 Generalized sign test

Existing two-sample median test for complete data is within the framework of linear rank test, and is the sign test procedure counting the number of observations exceeding the median of the pooled-sample. Let M_0, n_j and c_j be the observed pooled median in two-sample, sample size and the number of observations exceeding M_0 in the sample j respectively, $j = 1, 2$. Exact distribution of c_j under H_0 based on the conditional inference is the hypergeometric distribution as well as one based on the inference for the contingency table. Based on the normal approximation, sign test statistic in two-sample median is

$$T_{\mathrm{HS}} = 4 \sum_{j=1}^{2} n_j (c_j/n_j - 0.5)^2 \tag{2.1}$$

which has asymptotically chi-squared distribution with one degree of freedom (Hájek and Sidák, 1967). It is well known that T_{HS} is locally most powerful rank test against the alternative hypothesis with location shifts in the double exponential distribution and also it only has an efficiency of about 64% compared to the t-test in the normal distribution (Yanagawa, 1982).

Let $T_{ji} = \min(X_{ji}, C_{ji})$ be observed survival times in the mutually independent sample j $(i = 1, \ldots, n_j)$. Let X_{ji} be non-negative continuous random variables following a distribution $1 - S_j$, and C_{ji} be censored times following a common distribution independently from X_{ji} $(i = 1, \ldots, n_j,\ j = 1, 2)$. As an extension of the sign test for censored data, Brookmeyer and Crowley (1982b) considered to count the number of observations less than $\widehat{M}_0$ along with generalized signs, where $\widehat{M}_0$ is an estimate of pooled median survival time based on the weighted Kaplan-Meier estimate. The weighted Kaplan-Meier estimate is given by

$$\widehat{S}_0(t) = n^{-1}\{n_1\widehat{S}_1(t) + n_2\widehat{S}_2(t)\},$$

where $\widehat{S}_j$ and $n(= n_1+n_2)$ are the Kaplan-Meier estimate in the sample j $(j = 1, 2)$ and the sample size for the pooled-sample respectively. Unlike the Kaplan-Meier estimate based on the combined sample regardless of group, $\widehat{S}_0(t)$ is constructed so that it remains unaffected by the difference in censoring distributions between groups as possible. Brookmeyer and Crowley (1982b) considers $\widehat{S}_0(t)$ as a continuous function by using linear interpolation, and they found $\widehat{M}_0$ as an estimate of M_0 by

$$\widehat{M}_0 = L_0 + \frac{(0.5 - \widehat{S}_0(L_0))(U_0 - L_0)}{\widehat{S}_0(U_0) - \widehat{S}_0(L_0)}, \tag{2.2}$$

where L_0 is the largest observed death time with $\widehat{S}_0(t) > 0.5$, and U_0 is the smallest observed death time with $\widehat{S}_0(t) < 0.5$. The estimated survival probability $S_j(\widehat{M}_0)$ at time $\widehat{M}_0$ in the sample j, found by linear interpolation, is given by

$$\widehat{S}_j^{\mathrm{lin}}(\widehat{M}_0) = \widehat{S}_j(L_{0j}) + \frac{\{\widehat{S}_j(U_{0j}) - \widehat{S}_j(L_{0j})\}(\widehat{M}_0 - L_{0j})}{(U_{0j} - L_{0j})},$$

where L_{0j} and U_{0j} are two consecutive death times in the sample j with $L_{0j} \leq \widehat{M}_0 < U_{0j}$, and $\widehat{S}_j^{\mathrm{lin}}(\widehat{M}_0)$ is the estimated probability that a randomly selected individual in the sample j exceeds $\widehat{M}_0$ $(j = 1, 2)$. Their test procedure evaluates the difference between $\widehat{S}_j^{\mathrm{lin}}(\widehat{M}_0)$ and the expected survival probability of 0.5 under H_0, and the test statistic

$$T_{\mathrm{BC}} = n(\widehat{S}_1^{\mathrm{lin}}(\widehat{M}_0) - 0.5)^2/n_2\sigma_1^2 \tag{2.3}$$

has asymptotically chi-squared distribution with one degree of freedom under H_0, where $\widehat{\sigma}_j^2$ is given by

$$\widehat{\sigma}_j^2 = \frac{n - n_j}{n}\left\{\widehat{\mathrm{Var}}\{\widehat{S}_1^{\mathrm{lin}}(\widehat{M}_0)\} + \widehat{\mathrm{Var}}\{\widehat{S}_2^{\mathrm{lin}}(\widehat{M}_0)\}\right\},$$

and $\widehat{\mathrm{Var}}\{\widehat{S}_j^{\mathrm{lin}}(\widehat{M}_0)\}$ is given by

$$\begin{aligned}\widehat{\mathrm{Var}}\{\widehat{S}_j^{\mathrm{lin}}(\widehat{M}_0)\} &= \left[\widehat{S}_j(U_{0j})\left(\frac{\widehat{M}_0 - L_{0j}}{U_{0j} - L_{0j}}\right)\right]^2 \int_0^{U_{0j}} \frac{d\overline{\mathcal{N}}_j(t)}{\overline{\mathcal{Y}}_j(t)(\overline{\mathcal{Y}}_j(t) - d\overline{\mathcal{N}}_j(t))} \\ &+ \left\{\left[\widehat{S}_j(L_{0j})\left(\frac{U_{0j} - \widehat{M}_0}{U_{0j} - L_{0j}}\right)\right]^2 + \frac{2(\widehat{M}_0 - L_{0j})(U_{0j} - \widehat{M}_0)}{(U_{0j} - L_{0j})^2}\widehat{S}_j(L_{0j})\widehat{S}_j(U_{0j})\right\} \\ &\times \int_0^{L_{0j}} \frac{d\overline{\mathcal{N}}_j(t)}{\overline{\mathcal{Y}}_j(t)(\overline{\mathcal{Y}}_j(t) - d\overline{\mathcal{N}}_j(t))},\end{aligned}$$

which is the variance estimate of $\widehat{S}_j^{\mathrm{lin}}(\widehat{M}_0)$, $j = 1, 2$ (Klein and Moeschberger, 2003, p.232). Note that $\overline{\mathcal{N}}_j(t)$ and $\overline{\mathcal{Y}}_j(t)$ are $\overline{\mathcal{N}}_j(t) = \sum_{i=1}^{n_j} \mathcal{N}_{ji}(t)$ and $\overline{\mathcal{Y}}_j(t) = \sum_{i=1}^{n_j} \mathcal{Y}_{ji}(t)$ respectively. And, each of the pair $(\mathcal{N}_{ji}(t), \mathcal{Y}_{ji}(t))$ is usual counting process and at risk process to be used in survival analyses, and they can also be expressed as $\mathcal{N}_{ji}(t) = \mathbb{1}(T_{ji} \le t, T_{ji} \le C_{ji})$ and $\mathcal{Y}_{ji}(t) = \mathbb{1}(t \le T_{ji})$ respectively, where, $\mathbb{1}(\cdot)$ denotes the indicator function.

One of the formal problems in (2.3) is that it does not provide the same statistics value for the replacement between sample 1 and sample 2. To obtain a form providing invariance for that replacement, however, T_{BC} can be modified to

$$T'_{\mathrm{BC}} = (\widehat{S}_1^{\mathrm{lin}}(\widehat{M}_0) - 0.5)^2/\sigma_1^2 + (\widehat{S}_2^{\mathrm{lin}}(\widehat{M}_0) - 0.5)^2/\sigma_2^2$$

with the linear combination of competitive statistics. In fact, T'_{BC} is reduced to (2.1) for complete data.

Since the generalized sign test is based on the asymptotic property of the Kaplan-Meier estimate, low power as well as typical sign test and the validity in small sample size are issued.

2.3 Empirical likelihood ratio test

Naik-Nimbalkar and Rajarshi (1997) proposed an empirical likelihood ratio test for testing $\mathrm{H}_0^{\mathrm{m}}$. Using the same notation in Section 2.1, let $T_{ji} = \min(X_{ji}, C_{ji})$ be observed survival times in the mutually independent sample j ($i = 1, \ldots, n_j$, $j = 1, 2$). And, let M_0^* be the common median under $\mathrm{H}_0^{\mathrm{m}}$. The unconstrained log likelihood for the sample j can be expressed using the terms in counting process as follows:

$$\log L_j = \sum_{i=1}^{n_j} \left\{ d\overline{\mathcal{N}}_j(T_{ji}) \log(\lambda_j(T_{ji})) + (\overline{\mathcal{Y}}_j(T_{ji}) - d\overline{\mathcal{N}}_j(T_{ji})) \log(1 - \lambda_j(T_{ji})) \right\}, \qquad (2.4)$$

where $\lambda_j(T_{ji})$ is hazard ($j = 1, 2$). It is well known that the unconstrained log likelihood (2.4) is maximum at $\widehat{\lambda}_j(t) = d\overline{\mathcal{N}}_j(t)/\overline{\mathcal{Y}}_j(t)$. Let this unconstrained maximum value be denoted by $\log L_{\rm u}(\widehat{\lambda}_j(t)) = \sum_j \log L_j$. In order to obtain the empirical likelihood for M_0^*, we maximize $\sum_{j=1}^2 \log L_j$ subject to the following two constraints:

$$S_j(M_0^*) = \prod_{t<M_0^*} (1 - \lambda_j(t)) = 0.5, \ j = 1, 2.$$

By using the Lagrangian Multiplier method for the constrained maximization problem, one can obtain the estimate of the constrained hazard $\lambda_j^*(t)$, that is, $\lambda_j(t)$ under $\mathrm{H}_0^{\rm m}$, by differentiating the following function with respect to α_j and $\lambda_j(t)$ ($j = 1, 2$):

$$\sum_{j=1}^{2} \left\{ \log L_j - \alpha_j \left\{ \sum_{t<M_0^*} \log(1 - \lambda_j(t)) - \log(0.5) \right\} \right\}, \tag{2.5}$$

where α_j is Lagrangian parameter, $j = 1, 2$. The resulting $\widehat{\lambda}_j^*(t)$ is given by

$$\widehat{\lambda}_j^*(t) = \begin{cases} \dfrac{d\overline{\mathcal{N}}_j(t)}{\overline{\mathcal{Y}}_j(t) + \alpha_j}, & t < M_0^*, \\ \widehat{\lambda}_j(t), & t \geq M_0^*. \end{cases} \tag{2.6}$$

Thus the maximum log likelihood subject to those constraints can be calculated by substituting $\widehat{\lambda}_j^*(t)$ into (2.4), $j = 1, 2$. Let this constrained maximum value be denoted by $\log L_{\rm c}(\widehat{\lambda}_j^*(t)$. Naik-Nimbalkar and Rajarshi (1997) showed that the empirical profile likelihood ratio statistic

$$T_{\rm ELR} = 2 \log \left(\frac{L_{\rm u}(\widehat{\lambda}_j(t))}{L_{\rm c}(\widehat{\lambda}_j^*(t))} \right)$$

has asymptotically a chi-square distribution with one degree of freedom under $\mathrm{H}_0^{\rm m}$.

In their paper, no numerical solution was provided. Therefore, we have two problems on the estimations of M_0^* and α_j in (2.6), $j = 1, 2$. Here, we propose to estimate them based on the Newton-Raphson method. It is reasonable to consider that M_0^* would lie between two observed median survival times. Let $\widehat{M}_j = \inf\{t : \widehat{S}_j(t) < 0.5\}$, $j = 1, 2$, be the observed median survival times, and assume $\widehat{M}_1 < \widehat{M}_2$ as an illustration. In practical case, we may experience $\widehat{M}_1 = \widehat{M}_2$. In such case, one does not need to find $\widehat{M}_0^*$ and $\widehat{\alpha}_j$, $j = 1, 2$, since the common median survival time can be considered $\widehat{M}_0^*$ so that the constrained maximum log likelihood equal to the unconstrained maximum log likelihood (in this case, p-value can be 1.0 because $T_{\rm ELR} = 0$). For the case $\widehat{M}_1 \neq \widehat{M}_2$, since $\widehat{M}_0^*$

is expected to lie between $\widehat{M}_1$ and $\widehat{M}_2$, $\lambda_1^*(t)$ should be adjusted to have smaller hazard while $\lambda_2^*(t)$ should be adjusted to have larger hazard for $t < \widehat{M}_0^*$. In other words, $\widehat{\alpha}_1$ must be positive, and $\widehat{\alpha}_2$ must be negative in (2.6).

In order to find $\widehat{\alpha}_j$ efficiently, and to avoid possible convergence failures through the Newton-Raphson method, we assume $\alpha_j > 0,\ j = 1, 2$. Instead, we re-define $\widehat{\lambda}_j^*(t) = d\overline{\mathcal{N}}_j(t)/(\overline{\mathcal{Y}}_j(t) + \psi_j\alpha_j)$ for $t < M_0^*$ in (2.6), where ψ_j takes either 1 (for sample with $\min(\widehat{M}_1, \widehat{M}_2)$) or -1 (for sample with $\max(\widehat{M}_1, \widehat{M}_2)$). Next, we define $\alpha_j = \exp(\beta_j)$ as transformation of variables. Thus the function of β_j of interest and its derivative for fixed M_0^* are given as follows:

$$\begin{aligned}
\mathcal{F}(\beta_j) &= \sum_{t<M_0^*} \log\left(1 - \frac{d\overline{\mathcal{N}}_j(t)}{\overline{\mathcal{Y}}_j(t) + \psi_j \exp(\beta_j)}\right) - \log(0.5), \\
\mathcal{F}'(\beta_j) &= \frac{d\mathcal{F}(\beta_j)}{d\beta_j} = -\sum_{t<M_0^{\mathrm{m}}} \frac{d\overline{\mathcal{N}}_j(t)\psi_j \exp(\beta_j)}{\{\overline{\mathcal{Y}}_j(t) + \psi_j \exp(\beta_j)\}\{\overline{\mathcal{Y}}_j(t) + \psi_j \exp(\beta_j) - d\overline{\mathcal{N}}_j(t)\}}.
\end{aligned}$$

And, one can start the following iteration process with initial guess $\beta_j^{(0)}$ until a convergence condition is satisfied:

$$\beta_j^{(\eta)} = \beta_j^{(\eta-1)} - \frac{\mathcal{F}(\beta_j^{(\eta-1)})}{\mathcal{F}'(\beta_j^{(\eta-1)})}.$$

Finally, $\widehat{\alpha}_j$ can be obtained by the logarithmic transformation of $\widehat{\beta}_j$, $j = 1, 2$.

The remaining problem is the estimation of M_0^*. We assume that all distinct death times between $\min(\widehat{M}_1, \widehat{M}_2)$ and $\max(\widehat{M}_1, \widehat{M}_2)$ are candidates of M_0^*. Let $t_{(1)}, t_{(2)}, \ldots, t_{(\xi)}$ and $\widehat{\alpha}_{j(l)}$ be all death times between $\min(\widehat{M}_1, \widehat{M}_2)$ and $\max(\widehat{M}_1, \widehat{M}_2)$, and the Lagrangian parameters found by the Newton-Raphson method at a death time $t_{(l)},\ l = 1, \ldots, \xi$. And, let $\log L_{c(l)}$ be the constrained log likelihood at $t_{(l)}$ which can be calculated based on $\widehat{\alpha}_{j(l)},\ l = 1, \ldots, \xi$. Here, we can define the constrained maximum log likelihood as $\max(\log L_{c(1)}, \log L_{c(2)}, \ldots, \log L_{c(\xi)})$, and can consider a death time with the constrained maximum log likelihood as $\widehat{M}_0^*$. If no death time is existed between medians, then both medians may be considered candidates of M_0^*.

Once $\widehat{M}_0^*$ and the associated $\widehat{\alpha}_{j(\cdot)}$ are found, one can construct the constrained Kaplan-

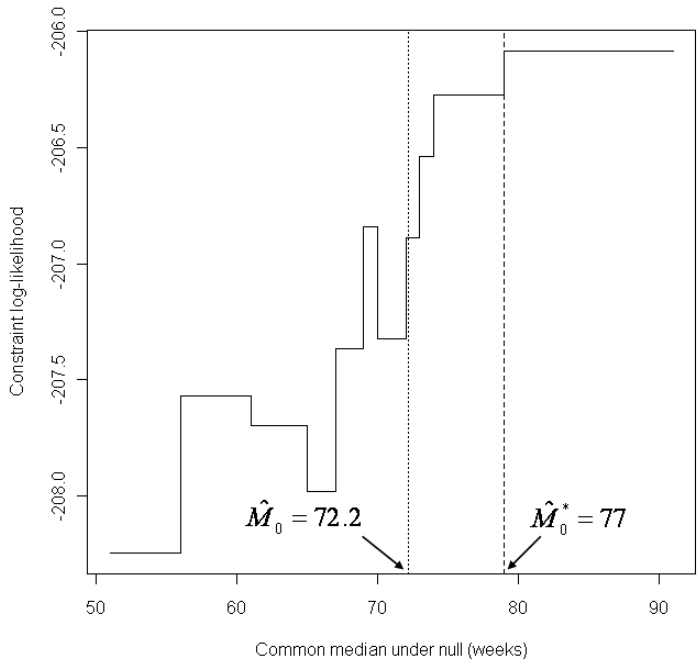

Figure 1: The behavior of the constrained log likelihoods for tongue cancer data

Meier estimate $\widehat{S}_j^{\mathrm{c}}$, $j = 1, 2$. The constrained Kaplan-Meier estimate is defined as

$$\widehat{S}_j^{\mathrm{c}}(t) = \begin{cases} \prod_{s \le t} \left(1 - \dfrac{d\overline{\mathcal{N}}_j(s)}{\overline{\mathcal{Y}}_j(s) + \psi_j \widehat{\alpha}_{j(\cdot)}} \right), & t < \widehat{M}_0^*, \\ \widehat{S}_j^{\mathrm{c}}(\widehat{M}_0^*) \prod_{\widehat{M}_0^* < s \le t} \left(1 - \dfrac{d\overline{\mathcal{N}}_j(s)}{\overline{\mathcal{Y}}_j(s)} \right), & t \ge \widehat{M}_0^*, \end{cases} \tag{2.7}$$

where we note that $\widehat{S}_1^{\mathrm{c}}(t)$ and $\widehat{S}_2^{\mathrm{c}}(t)$ have the estimate of common median $\widehat{M}_0^*$.

To illustrate the behavior of constrained log likelihoods between $\widehat{M}_1$ and $\widehat{M}_2$ graphically, we apply the empirical likelihood approach to survival data for patients with tongue cancer (Sickle-Santanello *et al.* 1988). This survival data was collected to investigate the effects of ploidy (aneuploid or diploid) on the prognosis of patients with tongue cancer. Here, let sample 1 and sample 2 be the aneuploid tumors group with 52 patients (21 censored observations) and the diploid tumors group with 28 patients (6) respectively. As a result, the estimates of median for each sample were $\widehat{M}_1 = 42$ (days) and $\widehat{M}_2 = 93$, and the common median under $\mathrm{H}_0^{\mathrm{m}}$ was estimated as $\widehat{M}_0^* = 77$. The estimates of Lagrangian parameters at $\widehat{M}_0^*$ providing the constrained maximum log likelihood were $\widehat{\alpha}_1 = 6.91$ and $\widehat{\alpha}_2 = 3.50$. Figure 1 shows the trend of the constrained log likelihoods with two time

points of $\widehat{M}_0^*$ and $\widehat{M}_0$ which is the estimate of the common median defined by (2.2). From this example, we found that $\widehat{M}_0^*$ is not the same as $\widehat{M}_0$, and the constrained log likelihoods does not show the convex shape. We also found that the constrained log likelihood changes at observed (non-censored) death times.

2.4 Bootstrap median test

Park and Na (2000) proposed a bootstrap median test for test $\mathrm{H}_0^{\mathrm{m}}$. Using the same notations in Section 2.1, let $T_{ji} = \min(X_{ji}, C_{ji})$ be observed survival times in the mutually independent sample j $(i = 1, \ldots, n_j,\ j = 1, 2)$. Thus, survival data can be conveniently represented by pairs of random variables (T_{ji}, δ_{ji}), where δ_{ji} indicates that lifetime X_{ji} is observed $(\delta_{ji} = 1)$ or (right-)censord $(\delta_{ji} = 0)$. Let $\widehat{M}_1$ and $\widehat{M}_2$ be observed medians for each sample, and x be the absolute difference of them, that is, $x = |\widehat{M}_1 - \widehat{M}_2|$. The bootstrap median test has the following steps:

1. Draw a sample with size of $n(= n_1 + n_2)$ by random re-sampling with replacement from the combined sample.

2. Allocate the first n_1 observations as $(T_{11}^{\mathrm{b}}, \delta_{11}^{\mathrm{b}}), \ldots, (T_{1n_1}^{\mathrm{b}}, \delta_{1n_1}^{\mathrm{b}})$ and the remaining n_2 observations as $(T_{21}^{\mathrm{b}}, \delta_{21}^{\mathrm{b}}), \ldots, (T_{2n_2}^{\mathrm{b}}, \delta_{2n_2}^{\mathrm{b}})$.

3. Calculate observed medians $\widehat{M}_1^{\mathrm{b}}$ and $\widehat{M}_2^{\mathrm{b}}$ for each of bootstrap samples. Define $x_{\mathrm{b}} = |\widehat{M}_1^{\mathrm{b}} - \widehat{M}_2^{\mathrm{b}}|$.

4. Repeat Step 1 to 3 B times.

For testing H_0, the approximate bootstrap p-value can be defined as the total number of x_{b} whose values are greater than or equal to x divided by B. Let

$$\widehat{p}_{boot} = \#\{x_{\mathrm{b}} \geq x\}/B.$$

The testing problem of $\mathrm{H}_0^{\mathrm{m}}$ is also referred in their paper (Park and Na, 2000). They consider using the difference of medians for this testing problem too, since one can compute p-value without a great deal amount of computing time. To obtain the null distribution, let $d = \widehat{M}_1 - \widehat{M}_2$. Unlike the testing problem of H_0, we don't need to combine two samples.

Instead, we need to do random re-sampling with replacement in each of samples where sample 2 consists of $(T_{21}+d, \delta_{21}), \ldots, (T_{2n_2}+d, \delta_{2n_2})$. By doing so, two medians for each sample can coincide. Thus, the approximate bootstrap p-value can be obtained with the same procedure Step 1 to 4.

2.5 Median test based on the order statistics

2.5.1 General theory

Let $X_{j(1)}, X_{j(2)}, \ldots, X_{j(n_j)}$ be n_j order statistics of random variables $X_{j1}, \ldots, X_{jn_j}$ with distribution $1 - S_j$, $j = 1, 2$. Let $f_{j(m_j)}(t)$ be the density function of $X_{j(m_j)}$ which is defined as

$$f_{j(m_j)}(t) = \{1 - S_j(t)\}^{m_j - 1}\{S_j(t)\}^{n_j - m_j} f_j(t)/\mathrm{B}(m_j, n_j - m_j + 1), \tag{2.8}$$

where $f_j(t)$ is the probability density function in the sample j, and $1/\mathrm{B}(m_j, n_j - m_j + 1) = n_j!/(m_j - 1)!(n_j - m_j)!$ is the binomial coefficient. Note that m_j can be the real number within the following discussions though it is usually considered as the integer. For simplicity, we discuss $f_{j(m_j)}(t)$ with the form of (2.8) even if m_j is not the integer. Thus, the function B is extended as the form of usual beta function.

As our interest is statistical inference for the difference in two-sample median survival times, suppose m_j is mid-order in the sample j, namely, $m_j = (n_j + 1)/2$. For complete data, median survival time is defined as

$$M_j = \begin{cases} \{X_{j(m_j - 1/2)} + X_{j(m_j + 1/2)}\}/2, & \text{if } n_j \text{ is even,} \\ X_{j(m_j)}, & \text{if } n_j \text{ is odd.} \end{cases} \quad (j = 1, 2) \tag{2.9}$$

Let random variables Y and V be

$$Y = M_1 - M_2, \quad V = M_1. \tag{2.10}$$

Thus, Y can be considered a random variable indicating the difference in two-sample median survival times. However, the definition of (Y, V) according to the case of even/odd of n_j in (2.10) through (2.9) poses very troublesome computations. Thus, we redefine Y and V as

$$Y = X_{1(m_1)} - X_{2(m_2)}, \quad V = X_{1(m_1)} \tag{2.11}$$

in order to avoid handling these random variables as in (2.10). By using the definition (2.11), the joint density of $(X_{1(m_1)}, X_{2(m_2)})$ is given by $\prod_{j=1}^{2} f_{j(m_j)}(t)$. And, the joint density function $g_1(y, v)$ of Y and V is defined as

$$g_1(y, v) = f_{1(m_1)}(v) f_{2(m_2)}(v - y)$$

by transformation of variables. The density function $g_1(y)$ of Y is defined as the marginal density function of $g_1(y, v)$, and given by

$$g_1(y) = B_{m_1 m_2} \int_{v \in [y, \infty)} \phi_1(v)\phi_2(v - y) f_2(v - y) f_1(v) dv, \tag{2.12}$$

where $B_{m_1 m_2} = 1/\mathrm{B}(m_1, m_1)\mathrm{B}(m_2, m_2)$, and ϕ_1 and ϕ_2 are

$$\phi_1(t) = S_1(t)^{m_1 - 1}\{1 - S_1(t)\}^{m_1 - 1}, \quad \phi_2(t) = S_2(t)^{m_2 - 1}\{1 - S_2(t)\}^{m_2 - 1},$$

respectively. So, $g(y)$ can be regarded as the density for the difference in 'two-sample mid-order statistics'. In other words, it can be referred as the density for the difference in two-sample median survival times. Let $x(= M_1 - M_2) \geq 0$ be the observed difference in two-sample median survival times, then a significance probability (p-value) function is defined as

$$pv_1(x) = \int_{y \in [x, \infty)} g_1(y) dy. \tag{2.13}$$

Equation (2.13) provides the probability for what the difference in two median survival times is greater than or equal to x. Meanwhile, let $pv_1^-(-x)$ be the probability below $-x$ and given by $pv_1^-(-x) = \int_{y \in (-\infty, -x]} g_1(y) dy$, then the value of $pv_1^-(-x)$ is the same as $pv_2(-x)$ which can be calculated via the same processes (2.12) and (2.13) under the replacement between sample 1 and 2 for g_1, where $g_2(y, v) = f_{2(m_2)}(v) f_{1(m_1)}(v - y)$. Thus, so called two sided significance probability can be calculated by $pv_1(x) + pv_2(-x)$. And, we can proceed to discussions with a form of equation (2.13) since the generality is not lost even if the equation is subject to $x \geq 0$.

We consider an asymptotic result for density function (2.12) below. Let $S_0(t)(= S_1(t) = S_2(t))$ and $M_0^* = S_0^{-1}(0.5)$ be true survival function and true median survival time under H_0, respectively. When $S_0(t)$ is twice differentiable function and H_0 is true, $g_1(y)$ is equivalent to the following normal density function asymptotically

$$\left\{2\pi V_0^*(n_1, n_2)\right\}^{-1/2} \exp\left\{-y^2 / 2V_0^*(n_1, n_2)\right\}, \tag{2.14}$$

where, $V_0^*(n_1, n_2)$ is expressed as

$$V_0^*(n_1, n_2) = \{1/4n_1 + 1/4n_2\}/f_0(M_0^*)^2.$$

Note that $f_0(t)$ is the probability density function of $S_0(t)$. This result can be shown based on either Laplace approximation to $g_1(y)$ (See Appendix A.1) or that the distribution of M_j under H_0 is equivalent to the normal distribution with mean M_0^* and variance $1/4n_j f_0(M_0^*)^2$ asymptotically (Desu and Raghavarao, 2004, p.152-153 for reference).

2.5.2 Discrete approximation of $pv_1(x)$

To utilize result of (2.14) for the testing problem, n_1 and n_2 are preferred to be large sufficiently. When n_1 and n_2 are not so large, we needs to consider the direct calculation of $pv_1(x)$ in the inference for the distribution of the difference in median survival times. However, it has drawbacks in computing $pv_1(x)$ defined by (2.13) since the computation of double integral is usually expected to be very complicated even if any parametric distribution is assumed. Actually, even if equation (2.13) is expanded exactly using the integration by parts, it causes very serious numeric errors since not only the digit varies a great deal depending on each of terms, but also a great number of positive and negative terms are also accrued. As a contrivance to avoid such errors, we propose a discrete approximation for (2.13). Such discrete approximation is useful for a case where S_1 and S_2 are replaced by discrete estimates such as Kaplan-Meier estimate.

Let $\tilde{S}_1$ and $\tilde{S}_2$ are approximations of S_1 and S_2 defined as step functions respectively, where the function jumps at $0 < t_1 < t_2 < \cdots < t_h$ (including zero jumps). Note that h is arbitrarily constant. In the approximation, the region of v is $v \in \{t_1, t_2, \ldots, t_h\}$, so we set $\tilde{v}$ as $v = t_{\tilde{v}}$, $\tilde{v} = 1, \ldots, h$. Using these notations, the discrete approximation of $g_1(y)dy$ in (2.13) is

$$\tilde{g}_1(y)dy = B_{m_1 m_2} \sum_{\{\tilde{v}: t_{\tilde{v}} \geq y\}} \tilde{\phi}_1(t_{\tilde{v}})\tilde{\phi}_2(t_{\tilde{v}} - y)d\tilde{S}_2(t_{\tilde{v}} - y)d\tilde{S}_1(t_{\tilde{v}}),$$

where $\tilde{\phi}_j(t) = \phi_j(t)|_{S_j(t)=\tilde{S}_j(t)}$, $j = 1, 2$. Let $\widetilde{pv}_1(x)(= \int_x^\infty \tilde{g}_1(y)dy)$ be an approximation

form of $pv_1(x)$. Based on the exchange order of integration, $\widetilde{pv}_1(x)$ can be expressed as

$$\begin{aligned}
\widetilde{pv}_1(x) &= \int_0^\infty \mathbb{1}(y \geq x) \left\{ B_{m_1 m_2} \sum_{\{\tilde{v}: t_{\tilde{v}} \geq y\}} \tilde{\phi}_1(t_{\tilde{v}}) \tilde{\phi}_2(t_{\tilde{v}} - y) d\tilde{S}_2(t_{\tilde{v}} - y) d\tilde{S}_1(t_{\tilde{v}}) \right\} \\
&= B_{m_1 m_2} \sum_{\tilde{v}=1}^{h} \int_0^\infty \mathbb{1}(y \geq x, t_{\tilde{v}} \geq y) \tilde{\phi}_1(t_{\tilde{v}}) \tilde{\phi}_2(t_{\tilde{v}} - y) d\tilde{S}_2(t_{\tilde{v}} - y) d\tilde{S}_1(t_{\tilde{v}}) \\
&= B_{m_1 m_2} \sum_{\tilde{v}=1}^{h} \tilde{\phi}_1(t_{\tilde{v}}) d\tilde{S}_1(t_{\tilde{v}}) \sum_{\{t_{\tilde{v}} \geq y \geq x\}} \tilde{\phi}_2(t_{\tilde{v}} - y) d\tilde{S}_2(t_{\tilde{v}} - y).
\end{aligned}$$

For simplicity, suppose y lies in range $t_{\bar{y}-1} < y \leq t_{\bar{y}}$, where $t_{\bar{y}}$ is the smallest time among $t_l \in \{t_1, t_2, \ldots, t_h\}$ with $t_{l-1} < y \leq t_l$. Then, the region of integration $\{\tilde{v} : t_{\tilde{v}} \geq y\}$ of $\tilde{g}_1(y)$ is equivalent to the set $\{t_{\bar{y}}, t_{\bar{y}+1}, t_{\bar{y}+2}, \cdots, t_h\}$. By transforming a variable to $y' = t_{\tilde{v}} - y$, the region of integration $\{x \leq y \leq t_{\tilde{v}}\}$ is transformed into the new region $\{0 \leq y' \leq t_{\tilde{v}} - x\}$, thus

$$\begin{aligned}
\widetilde{pv}_1(x) &= B_{m_1 m_2} \sum_{\tilde{v}=1}^{h} \tilde{\phi}_1(t_{\tilde{v}}) d\tilde{S}_1(t_{\tilde{v}}) \int_{\{0 \leq y' \leq t_{\tilde{v}} - x\}} \tilde{\phi}_2(y') d\tilde{S}_2(y') \\
&= B_{m_1 m_2} \sum_{\tilde{v}=1}^{h} \tilde{\phi}_1(t_{\tilde{v}}) d\tilde{S}_1(t_{\tilde{v}}) \sum_{l=1}^{\tilde{v}} \mathbb{1}(0 \leq t_l \leq t_{\tilde{v}} - x) \tilde{\phi}_2(t_l) d\tilde{S}_2(t_l) \qquad (2.15)
\end{aligned}$$

Note that $\sup_l |t_l - t_{l-1}| \to 0$ as $h \to \infty$, so that $\widetilde{pv}_1(x) \to pv_1(x)$.

2.5.3 Impact of censored data

In this section, we consider how general theory of the inference for the difference in median survival times based on order statistics from complete data can be applied to censored data. For censored data, median survival time can not be observed based on (2.9) typically. One of the definitions which is frequently used for the estimation of median survival time, $\inf\{t : \widehat{S}_j(t) < 0.5\}$, is not equal to (2.9) exactly in complete data. Similar definition to (2.2) has also the same problem. So, we define the observed (estimated) median survival time for censored data in the sample j as follows, so that it can be reduced to (2.9) in complete data

$$\widehat{M}_j = \begin{cases} U_j, & \widehat{S}_j^{\mathrm{m}} \leq 0.5, \\ U_j + (U_j - L_j)(0.5 - \widehat{S}_j^{\mathrm{m}}) / (\widehat{S}_j(U_j) - \widehat{S}_j(L_j)), & \widehat{S}_j^{\mathrm{m}} > 0.5, \end{cases} \qquad (2.16)$$

where $\widehat{S}_j^{\mathrm{m}} = \{\widehat{S}_j(L_j) + \widehat{S}_j(U_j)\}/2$, and L_j (U_j) is the largest (smallest) observed death time with $\widehat{S}_j(t) > 0.5$ ($\widehat{S}_j(t) < 0.5$) in the sample j, $j = 1, 2$.

Let $\widehat{g}_1(y)$ be the density function for the difference in median survival times, $y = \widehat{M}_1 - \widehat{M}_2$, in censored data. For censored data, note that the distribution of $\widehat{M}_j$ is asymptotically equivalent to the normal distribution with mean M_0^* and variance $\mathrm{Var}(\widehat{S}_j(M_0^*))/f_0(M_0^*)^2$ under H_0 when $S_0(t)$ is twice differentiable (Wang and Hettmansperger, 1990). Thus, based on the similar inference drawing (2.14), $\widehat{g}(y)$ is asymptotically equivalent to the following normal density function

$$\{2\pi\widehat{V}_0^*(n_1, n_2)\}^{-1/2} \exp\left\{-y^2/2\widehat{V}_0^*(n_1, n_2)\right\}. \tag{2.17}$$

In particular, the form of variance function $\widehat{V}_0^*(n_1, n_2)$ is given by

$$\widehat{V}_0^*(n_1, n_2) = \{\mathrm{Var}(\widehat{S}_1(M_0^*)) + \mathrm{Var}(\widehat{S}_2(M_0^*))\}/f_0(M_0^*)^2,$$

where the incompleteness due to censoring from $V_0^*(n_1, n_2)$ in complete data is considered. This result yields that the quantity $\widehat{M}_1 - \widehat{M}_2$ is orthogonally decomposed into $M_1 - M_2 \oplus Z_{\mathrm{cens}}$ approximately, and the normal univariate Z_{cens} has mean zero and variance $\widehat{V}_0^*(n_1, n_2) - V_0^*(n_1, n_2)$. To consider this result from a viewpoint of sample size, suppose (n_1', n_2') are sample sizes of each sample in complete data which are associated with original sample sizes (n_1, n_2) in censored data. Sample sizes (n_1', n_2') satisfies the following relationship to have the same quantity of test statistic based on (n_1, n_2)

$$\frac{\widehat{V}_0^*(n_1, n_2)}{\widehat{V}_0^*(n_1', n_2')} = \frac{\mathrm{Var}(\widehat{S}_1(M_0^*)) + \mathrm{Var}(\widehat{S}_2(M_0^*))}{1/4n_1' + 1/4n_2'} = 1.$$

Although there are a few methods to resolve above equation, we assume that current ratio of sizes $\gamma_{21} = n_2/n_1$ is fixed, thus

$$n_1' = (1 + 1/\gamma_{21})/4\{\mathrm{Var}(\widehat{S}_1(M_0^*)) + \mathrm{Var}(\widehat{S}_2(M_0^*))\}, \;\; n_2' = \gamma_{21} n_1'. \tag{2.18}$$

As a result, we need to compute (2.13) or (2.15) by substituting (n_1', n_2') for original sample sizes (n_1, n_2) when we consider using $pv_1(x)$ for testing problem of H_0 under censored data. While there are some considerable estimates to substitute the quantity $\mathrm{Var}(\widehat{S}_1(M_0^*)) + \mathrm{Var}(\widehat{S}_2(M_0^*))$ in (2.18), we use an estimate $\widehat{\mathrm{Var}}(\log \widehat{S}_1(\widehat{M}_1))/4 + \widehat{\mathrm{Var}}(\log \widehat{S}_2(\widehat{M}_2))/4$. Here, we use the Nelson-Aalen estimate $\int_0^{\widehat{M}_j} d\overline{\mathcal{N}}_j(t)/\overline{\mathcal{Y}}_j(t)(\overline{\mathcal{Y}}_j(t) - \overline{\mathcal{N}}_j(t))$ at a time-point $\widehat{M}_j$ for $\widehat{\mathrm{Var}}(\log \widehat{S}_j(\widehat{M}_j))$, $j = 1, 2$. An approximation $\mathrm{Var}(\widehat{S}_j(t)) \approx S_j(t)^2 \mathrm{Var}(\log \widehat{S}_j(t))$ in the derivation of Greenwood formula and $S_j(M_0^*) = 1/2$ are behind the use of such estimate.

2.5.4 Testing problem and the estimate of $pv_1(x)$

We discuss the testing problem of H_0 under situation where n_1 and n_2 are not sufficiently large. That is, we go over the use of $pv_1(x)$ or $\widetilde{pv}_1(x)$ as an alternative since we can't use the asymptotic result (2.14) immediately. However, $pv_1(x)$ has to be estimated by using some methods because S_1 and S_2 are unknown in actual survival data. An issue is how to construct better estimate of $pv_1(x)$. One of the solutions is the permutation procedure, but it causes serious computational load. Another procedure is to estimate $pv_1(x)$ by substituting the Kaplan-Meier estimates for S_1 and S_2, as well as V-statistic. Hence, we discuss the property of estimate and the associated problems when the Kaplan-Meier estimate is substituted into $\widetilde{pv}_1(x)$ under H_0. Because of $S_1 = S_2$ under H_0, we use the following weighted Kaplan-Meier estimate as the estimate of the true survival function $S_0(t)$ which is used in the testing procedure by Brookmeyer and Crowley (1982b)

$$\widehat{S}_0(t) = n^{-1}\{n_1\widehat{S}_1(t) + n_2\widehat{S}_2(t)\}.$$

And, we define $\widehat{pv}_1(x)$ which use $\widehat{S}_0$ as a substitute for $\tilde{S}_1$ and $\tilde{S}_2$ in (2.15). Then, $\widehat{pv}_1(x)$ is given by

$$\widehat{pv}_1(x) = B_{m_1m_2}\sum_{\tilde{v}=1}^{d}\widehat{\phi}_1(t_{\tilde{v}-})d\widehat{S}_0(t_{\tilde{v}})\sum_{l=1}^{\tilde{v}}\mathbb{1}(0 \le t_l \le t_{\tilde{v}} - x)\widehat{\phi}_2(t_{l-})d\widehat{S}_0(t_l) \quad (x \ge 0)$$

Here, $\{t_1, t_2, \cdots, t_d\}$ are the distinct observed death times ($h = d$) in the pooled sample, and $\widehat{\phi}_j(t-)$ is given by

$$\widehat{\phi}_j(t_-) = \widehat{S}_0(t_-)^{m_j-1}\{1 - \widehat{S}_0(t_-)\}^{m_j-1} \quad (j = 1, 2)$$

where t_- is a time just prior to t.

First we state the outline of asymptotic results for $\widehat{pv}_1(x)$. $\widehat{pv}_1(x)$ can be expressed using the terms used in the counting process:

$$\widehat{pv}_1(x) = B_{m_1m_2}\int_x^{\infty}\widehat{\phi}_1(v_-)\widehat{S}_0(v_-)\left\{\int_0^{v-x}\widehat{\phi}_2(t_-)\widehat{S}_0(t_-)\frac{d\overline{\mathcal{N}}(t)}{\overline{\mathcal{Y}}(t)}\right\}\frac{d\overline{\mathcal{N}}(v)}{\overline{\mathcal{Y}}(v)},$$

where $d\widehat{S}_0(t) = \widehat{S}_0(t_-)d\overline{\mathcal{N}}(t)/\overline{\mathcal{Y}}(t)$, $\overline{\mathcal{N}}(t) = \sum_{j=1}^{2}\overline{\mathcal{N}}_j(t)$ and $\overline{\mathcal{Y}}(t) = \sum_{j=1}^{2}\overline{\mathcal{Y}}_j(t)$. The Doob-Meyer decomposition of $\overline{\mathcal{N}}(t)$ under H_0 can be written as $\overline{\mathcal{N}}(t) = \overline{\mathcal{M}}(t) + \int_0^t \overline{\mathcal{Y}}(s)d\Lambda_0(s)$

(Fleming and Harrington, 1991; Andersen *et.al*, 1993). Here, $\Lambda_0(t)$ is the cumulative hazard function $-\log(S_0(t))$, and $\overline{\mathcal{M}}(t) = \sum_{i=1}^{n} \mathcal{M}_i(t)$ is $\mathscr{F}_t$-martingale process. Filtration $\mathscr{F}_t$ is the history just before t

$$\mathscr{F}_t = \sigma\{\mathcal{N}_{ji}(s), \mathcal{Y}_{ji}(s_+),\ 0 \le s \le t,\ i = 1, \cdots, n_j,\ j = 1, 2, \}$$

where t_+ is the value just after t. $\widehat{pv}_1(x)$ can be expressed by applying the Doob-Meyer decomposition to $\overline{\mathcal{N}}(t)$ as follows:

$$\begin{aligned} \widehat{pv}_1(x) &= B_{m_1 m_2} \textstyle\int_x^{\tau_n} \widetilde{\mathcal{M}}_2(v-x) d\widetilde{\mathcal{M}}_1(v) + B_{m_1 m_2} \int_x^{\tau_n} \widetilde{\mathcal{A}}_2(v-x) d\widetilde{\mathcal{M}}_1(v) \\ &\quad + B_{m_1 m_2} \textstyle\int_x^{\tau_n} \widetilde{\mathcal{M}}_2(v-x) d\widetilde{\mathcal{A}}_1(v) + B_{m_1 m_2} \int_x^{\tau} \widetilde{\mathcal{A}}_2(v-x) d\widetilde{\mathcal{A}}_1(v). \quad (2.19) \end{aligned}$$

Here, since $\widehat{\phi}_j(t_-)\widehat{S}_0(t_-)$ is $\mathscr{F}_t$-predictable, therefore, based on the martingale transform, $\widetilde{\mathcal{M}}_j(v) = \int_0^v \widehat{\phi}_j(t_-)\widehat{S}_0(t_-) d\overline{\mathcal{M}}(t)/\overline{\mathcal{Y}}(t)$ is $\mathscr{F}_v$-martingale process, and $\widetilde{\mathcal{A}}_j(v) = \int_0^v \widehat{\phi}_j(t_-)\widehat{S}_0$ $(t_-) d\Lambda_0(t)$ is $\mathscr{F}_v$-predictable process, $j = 1, 2$. Furthermore, $\tau_n = \sup\{t : \overline{\mathcal{Y}}(t) > 0\}$ indicates the maximum observation, and $\tau = \sup\{t : \mathrm{E}[\mathcal{Y}_{ij}(t)] = \Pr(t \le T_{ij}) > 0\}$ is limit of τ_n. For the later discussions, let $\widehat{pv}_{1j}(x)$ be jth term of four decomposed terms in (2.19) $(j = 1, \cdots, 4)$. Here, because of $\widehat{S}_0(t) = 0$ for $\tau_n \le t$, $\widehat{pv}_{14}(x)$ provides the same value even if the upper limit of integration τ is replaced by τ_n or ∞. So, we set τ beforehand for the upper limit of integral range in (2.19). However, we set τ_n for the upper limit of integral range in remaining three terms because it is necessary to consider it in the calculation of the predictable variation process.

First, we provide one of the asymptotic properties of $\widehat{pv}_1(x)$ as follows.

Theorem 1. For two-sample with sizes (n_1, n_2), let $\max(\widehat{M}_1, \widehat{M}_2) < \tau_n$ when H_0 is true. Also, assume that $S_0(t)$ is twice differentiable in the neighborhood $(\ni t)$ of M_0^*, and $\lim_{n\to\infty} \overline{\mathcal{Y}}(\tau_-) \to_p \infty$. Let $\min(n_1, n_2) \to \infty$, then $|\widehat{pv}_1(x) - pv_1(x)| \to_p 0$ on $x \in [0, \tau]$ (on the scale of $x = O(\sqrt{n/n_1 n_2})$ uniformly, where "$\to_p$" represents convergence in probability.

Proof of Theorem 1. To clarify the dependences of both $\widehat{pv}_1(x)$ and $pv_1(x)$ to m_1, m_2, $\widehat{S}_0$ and S_0, let $\widehat{pv}_1(x) = pv_1(x)^{(m_j, \widehat{S}_0^{(n)})}$. Using the same notation, $pv_1(x) = pv_1(x)^{(m_j, S_0)}$. Although we can make similar discussions given in Appendix A.1 in order to achieve our intended result, we show that each of two terms in right side given below convergence to

zero in probability uniformly on x:

$$|pv_1(x)^{(m_j,\widehat{S}_0^{(n)})} - pv_1(x)^{(m_j,S_0)}| \leq |pv_1(x)^{(m_j,\widehat{S}_0^{(n)})} - pv_1^*(x)^{(m_j,S_0)}| + |pv_1^*(x)^{(m_j,S_0)} - pv_1(x)^{(m_j,S_0)}|,$$

where $pv_1^*(x)^{(m_j,S_0)}$ is

$$pv_1^*(x)^{(m_j,S_0)} = B_{m_1 m_2} \int_x^\tau \left\{ \int_0^{v-x} \phi_1(v) S_0(v) \phi_2(t) S_0(t) d\Lambda_0(t) \right\} d\Lambda_0(v) \ \ (\leq pv_1(x)).$$

We use $pv_1^*(x)$ as simplified notation of $pv_1^*(x)^{(m_j,S_0)}$ as well as $\widehat{pv}_1(x)$ and $pv_1(x)$.

First, we assume that m_1 and m_2 are fixed, that is, $B_{m_1 m_2}$ and the power m_j in $\widehat{\phi}_j(\cdot)$ are fixed respectively, as a matter of form. Then, we show $\widehat{pv}_1(x)^{(m_j,\widehat{S}_0^{(n)})} \to_p pv_1^*(x)^{(m_j,S_0)}$ as $n \to \infty$. The consistency of Kaplan-Meier estimate $\widehat{S}_0$ can be obtained by a condition $\lim_{n\to\infty} \overline{\mathcal{Y}}(\tau_-) \to_p \infty$ (Fleming and Harrington, 1991, Theorem 3.4.2). Thus, based on that $\widehat{\phi}_j(v)\widehat{S}_0(v)\widehat{\phi}_l(t)\widehat{S}_0(v) \to_p \phi_j(v)S_0(v)\phi_l(v)S_0(v)$ uniformly on $v, t \in [0,\tau]$ $(j,l = 1,2)$ by the Continuous Mapping Theorem, and the application of Lenglart's Inequality (Andersen and Gill, 1982, Theorem I.1) to martingale components in $\widehat{pv}_{1j}(x)$ $(j = 1,2,3)$, we can show $\widehat{pv}_1(x) \to_p pv_1^*(x)$ uniformly on $x \in [0,\tau]$. Illustrating the application here, the predictive variation process of $\widetilde{\mathcal{M}}_j(v)$ in (2.19) is $\langle \widetilde{\mathcal{M}}_j \rangle(v) = \int_0^v \widehat{\phi}_j(t_-)^2 \widehat{S}_0(t_-)^2 d\Lambda_0(t)/\overline{\mathcal{Y}}(t)$, so $n\langle \widetilde{\mathcal{M}}_j \rangle(\tau) \to_p O(1)\mathrm{B}(2m_j - 1, 2m_j - 1)$ as $n \to \infty$. And, we obtain $\sup_{v\in[0,\tau]} |\widetilde{\mathcal{M}}_j(v)| \leq O_p(1)\{\mathrm{B}(2m_j - 1, 2m_j - 1)/n\}^{1/2}$ by Lenglart's Inequality. Through these discussions, we get that $\widehat{pv}_{1j}(x)$ converges to zero in probability, $j = 1,2,3$.

Next, we discuss $|pv_1^*(x)^{(m_j,S_0)} - pv_1(x)^{(m_j,S_0)}| \to_p 0$ as $m_1, m_2 \to \infty$ well as n. This can be shown by similar discussions in Appendix A.1, where $S_0(t)$ is twice differentiable. However, $\frac{d}{dx} pv_1(x)$ must be point density function when on the original scale of x when $m_1, m_2 \to \infty$. We avoid this by scale transform to $x \mapsto x\sqrt{n_1 n_2/n}$, then we apply the results of Laplace approximation (Appendix A.1). Here, since it is ensured that τ exceeds the true median asymptotically by a condition $\max(\widehat{M}_1, \widehat{M}_2) < \tau_n \leq \tau$, the results of Laplace approximation which implies that $pv_1^*(x)$ is dominated by the integrand around median, concludes that $pv_1^*(x)$ is identical with $pv_1(x)$ asymptotically. That is, we obtain $|pv_1^*(x) - pv_1(x)| \to_p 0$ uniformly on $x = O(\sqrt{n/n_1 n_2})$. Finally, we achieve our intended result by integrating these two results with Fleming and Harrington(1991, p.220) or Sugimoto and Goto(2003, Appendix A.3). □

From the relationship $pv_1^*(x) \leq pv_1(x)$, it seems to have a problem of integration regions in using $\widehat{pv}_1(x)$. For example, it is commonly occurred that τ_n and τ are not

close enough to $\inf\{t : S_0(t) = 0\}$, especially, when any fixed censoring time is happened. However, as well as Theorem 1., such problem can be negligible because $\widehat{pv}_1(x)$ is dominated by the integrand value around median when we have adequate sample sizes. Further investigations can be done based on simulation study, but we consider that trend what $\widehat{pv}_1(x)$ underestimates $pv_1(x)$ due to integration region for small samples is almost negligible practically.

Now, we investigate the relationship between $\mathrm{E}[\widehat{pv}_1(x)]$ and $pv_1^*(x)$ as one of properties in the finite sample. $\widehat{pv}_{11}(x)$ and $\widehat{pv}_{12}(x)$ holds martingale property for $\{\mathscr{F}_t, t \geq 0\}$ because each of integrands $\widetilde{\mathcal{M}}_2(v-x) \mathrel{\wr} \widetilde{\mathcal{A}}_2(v-x)$ is $\mathscr{F}_v$-predictable, so that their expectations are zero. Meanwhile, although the third term does not hold martingale property in the finite sample, it can be shown that $\widehat{pv}_{13}(x) \to_p 0$ by Theorem 1. and $\mathrm{E}[\widehat{pv}_{13}(x)] = o(n^{-1})$ by discussions in Appendix A.2. Therefore, expectation of $\widehat{pv}_1(x)$ can be written as $\mathrm{E}[\widehat{pv}_1(x)] = \mathrm{E}[\widehat{pv}_{14}(x)] + o(n^{-1})$, and

$$\mathrm{E}[\widehat{pv}_{14}(x)] = B_{m_1 m_2} \int_x^\tau \int_0^{v-x} \mathrm{E}\left[\widehat{\phi}_1(v_-)\widehat{S}_0(v_-)\widehat{\phi}_2(t_-)\widehat{S}_0(t_-)\right] d\Lambda_0(t) d\Lambda_0(v). \quad (2.20)$$

In the discussion about the relationship between $\mathrm{E}[\widehat{pv}_{14}(x)]$ and $pv_1^*(x)$, we note that $\widehat{S}_0(t)$ has positive bias with the following range:

$$0 \leq \mathrm{E}[\widehat{S}_0(t)] - S_0(t) \leq (1 - S_0(t)) \exp\left(-\mathrm{E}[\overline{\mathcal{Y}}(t)]\right) \quad (2.21)$$

(Andersen *et al.*, 1993, p.259). For instance, for any functions $\psi_v(t)$ and $\psi_t(v)$, we have

$$\begin{aligned} &\mathrm{E}\left[\widehat{\phi}_1(v)\widehat{S}_0(v)\widehat{\phi}_2(t)\widehat{S}_0(t)\right] - \phi_1(v)S_0(v)\phi_2(t)S_0(t) \\ &\qquad = \psi_v(t)(\mathrm{E}[\widehat{S}_0(t)] - S_0(t)) + \psi_t(v)(\mathrm{E}[\widehat{S}_0(v)] - S_0(v)) \end{aligned} \quad (2.22)$$

by Taylor expansion of $\widehat{\phi}_1(v)\widehat{S}_0(v)\widehat{\phi}_2(t)\widehat{S}_0(t)$ around the true value and the mean value theorem. Here, $\psi_u(s)$ tends to have negative values for region $\{s \leq M_0^*\}$, and positive values for region $\{M_0^* \leq s\}$ $(u, s = v, t)$. By (2.21), we know that bias of $\mathrm{E}[\widehat{S}_0(s)] - S_0(s)$ in region $\{M_0^* \leq s\}$, in general, is bigger than one in region $\{s \leq M_0^*\}$ $(s = v, t)$. So, bias written by (2.22) has more regions of positive value than that of negative value on (t, v), so that $\mathrm{E}[\widehat{pv}_{14}(x)] - pv_1^*(x)$ has positive trend as a result of integration. Summarizing above discussions, $\mathrm{E}[\widehat{pv}_1(x)] - pv_1(x)$ is positive-biased slightly for small samples, and which implies that the procedure using $\widehat{pv}_1(x)$ provides a conservative result practically for testing H_0.

2.5.5 A few considerations for the computational problem

The calculation of proposed median test in two samples can be summarized as below. With respect to the definitional identity (2.16) of $\widehat{M}_j$, observed difference in medians x is defined by $x = \widehat{M}_1 - \widehat{M}_2$. The value of m_j in $\widehat{pv}_1(|x|)$ is $m_j = (n_j + 1)/2$ for complete data, however, we set $m_j = (n'_j + 1)/2$ with (n'_1, n'_2) which can be found by (2.18) for censored data ($j = 1, 2$). And, let $\widehat{pv}_2(|x|)$ be the one defined by the replacement of information between sample 1 and 2 in $\widehat{pv}_1(|x|)$. Under these settings, the two-sided significance probability is $\widehat{pv}_1(|x|) + \widehat{pv}_2(|x|)$ for testing problem of H_0. Hence, we define the two-sided significance probability as $\widehat{pv}(x) = \widehat{pv}_1(|x|) + \widehat{pv}_2(|x|)$.

One problem under the calculation process is that the value of $\widehat{pv}_1(0) + \widehat{pv}_2(0+)$ does not always become 1.0 exactly since $\widehat{pv}_j(x)$ is the discrete approximation of $pv_j(x)$. Also, another problem may be caused by the value of $B_{m_1 m_2}$ in large sample sizes. When n_j and m_j ($j = 1, 2$) are large, $B_{m_1 m_2}$ has explosively large value while $B^{-1}_{m_1 m_2}\widehat{pv}_j(|x|)$ has extremely small value. Here, as a result of loss of significant digits caused by both, the value of $\widehat{pv}_1(0)+\widehat{pv}_2(0+)$ calculated with $B_{m_1 m_2}$ is away from 1.0 further. Under situation where $n \leq 200$ for complete data, we observed that $\widehat{pv}_1(0)+\widehat{pv}_2(0+)$ is nearly equal to 1.0 even the discrete approximation so far. However, we also observed that its discrepancy from 1.0 becomes considerably serious as n becomes larger than 200. But, to overcome such computational problem, we can avoid such process as the computation of $\widehat{pv}_j(|x|)$ based on the direct calculation of $B_{m_1 m_2}$. That is, we assume $\widehat{pv}_1(0) + \widehat{pv}_2(0+) = 1$ on a constant basis, and employ the procedure which calculate the value of $\widehat{pv}(x)$ with the following form

$$B^{-1}_{m_1 m_2}\widehat{pv}(x)/\{B^{-1}_{m_1 m_2}(\widehat{pv}_1(0) + \widehat{pv}_2(0+))\} \tag{2.23}$$

where both denominator and the numerator have been divided by $B_{m_1 m_2}$ preliminarily.

Computational method by (2.23) has a merit not only to obtain the stable $\widehat{pv}(x)$ numerically but also to reduce the positive bias slightly inherent in $E[\widehat{pv}(x)]$. Further investigations for the property of $\widehat{pv}(x)$ by (2.23) in the finite sample are referred in the simulation study.

2.5.6 Testing problem for $\mathrm{H}_0^{\mathrm{m}}$

Now, we discuss testing problem for $\mathrm{H}_0^{\mathrm{m}}$. Although, our main concern may be the testing problem for H_0, it is valuable to resolve testing problem for $\mathrm{H}_0^{\mathrm{m}}$. To estimate $pv_1(x)$ under $\mathrm{H}_0^{\mathrm{m}}$ where two survival distributions have the common median survival time $M_0^*(= M_1^* = M_2^*)$, we propose to use the constrained Kaplan-Meier estimate $\widehat{S}_j^{\mathrm{c}}(t)$ $(j = 1, 2)$ given by (2.7). When the empirical likelihood ratio test is applied, $\widehat{S}_j^{\mathrm{c}}(t)$ $(j = 1, 2)$ are constructed based on the Lagrangian Multiplier method. And, the estimate $\widehat{M}_0^*$ having the maximum profile likelihood is also found there. Thus, we obtain $pv_1(x)$ under $\mathrm{H}_0^{\mathrm{m}}$ by substituting $\widehat{S}_j^{\mathrm{c}}(t)$ $(j = 1, 2)$ into $\widetilde{pv}_1(x)$.

3. Simulation Study

Two simulation studies are carried out to investigate the performance of existing median tests. First simulation study aims to compare the standard median test developed by Brookmeyer and Crowley (1982b) to proposed median test based on the property of order statistics, with respect to the shape of null distribution, empirical type I error and power under a simple simulation model. It is not easy to carry out the simulation study for many scenarios of alternative hypothesis, therefore we take a typical simulation model first for which the discrepancy between two survival distributions becomes large over time. Second simulation study aims to investigate the performance with respect to the shape of null distribution, empirical type I error and power for each of four median tests introduced in Section 2 under shift model. Unlike first simulation model, we focus on only shift alternative here since the shift model has been frequently used in the simulation for the nonparametric approach. Another reason is that median test is locally powerful against location shifts with censored data (Brookmeyer and Crowley, 1982b). In this simulation study based on the shift model, several underlying survival distributions are assumed.

3.1 Simulation study 1

Simulation Model: Assume that distribution function $1 - S_1(t)$ is uniform on $[0, 500]$. And, assume that survival function $S_2(t)$ is also a linear function with the same slope as $S_1(t)$ till a branch time-point $\xi(\geq t)$ and different slope after ξ. Thus, survival functions for each sample are defined as

$$S_1(t) = 1 - \frac{t}{500}, \qquad S_2(t) = \begin{cases} S_1(t), & t \leq \xi \\ b_2 + b_1(t - \xi), & t > \xi \end{cases}$$

where b_1, b_2 and ξ are

$$b_1 = a/(\delta - 500a), \quad b_2 = 0.5 + a, \quad \xi = 250 - 500a \quad (a > 0)$$

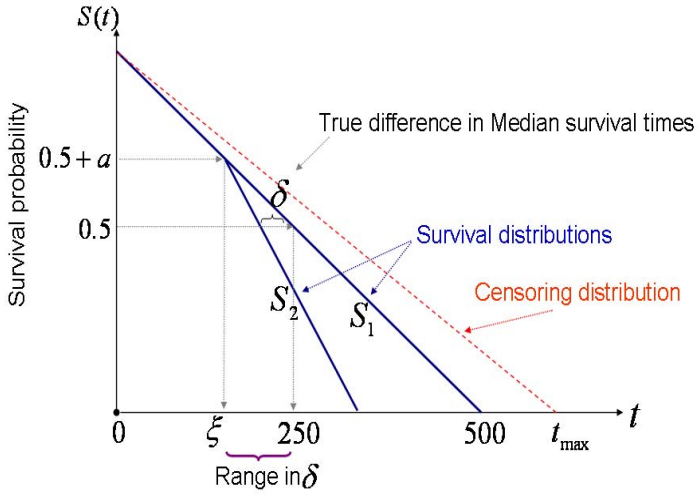

Figure 2: Graphical illustration of simulation

for a constant a (equally, survival probability $0.5+a$ at a branch time-point ξ) and a true difference in median survival times $\delta(= M_2^* - M_1^*)$ in Figure 2.

Censoring times C_{ji}, $i = 1, \cdots, n$, $j = 1, 2$ are generated from the uniform distribution on $[0, t_{\max}]$. In the simulation, four levels of the percentage of censoring, namely, 0%, 10%, 30% and 50% are considered respectively. The percentage of censoring is adjusted by moving $t_{\max}$ so that the targeted percentage of censoring is met. We set equal sample size for each of two samples, namely, $n_1 = n_2$ =15, 25, 50 and 100 are considered respectively.

3.1.1 Comparison of null distributions

In section 2.5.4, we discussed the approximate property of $\widehat{pv}_1(|x|)$ under H_0 and some trends of $E[\widehat{pv}_1(|x|)]$ in the finite sample, and then we provided a practical calculation method for both sided significance probability function $\widehat{pv}(x)$ in section 2.5.5. Here, we investigate the behavior of the null distribution of $\widehat{pv}(x)$ based on the simulation study in order to complement those discussions and to check the performance of $\widehat{pv}(x)$ in the finite sample.

For simulated data under H_0, we find p-values obtained by both $\widehat{pv}(x)$ and T'_{BC} which is the modified test statistic of testing procedure by Brookmeyer and Crowley (1982b). As

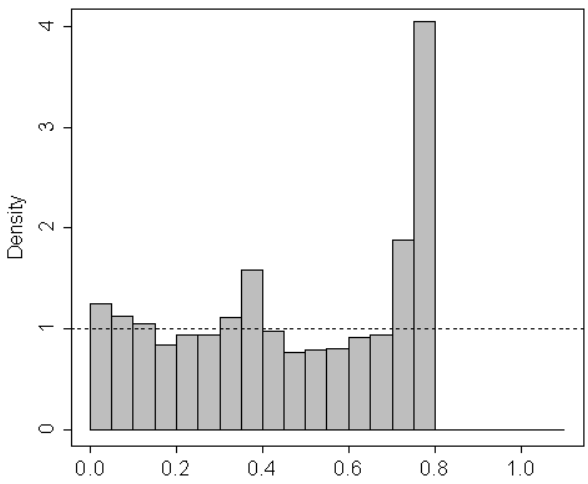

Figure 3: p-values (T'_{BC}) when $n = 30$

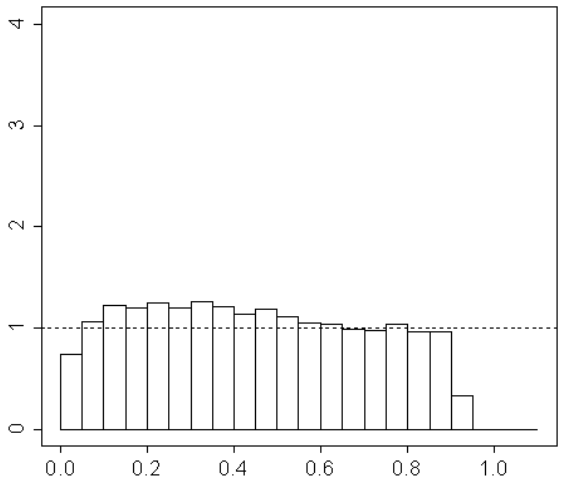

Figure 4: p-values ($\widehat{pv}(x)$) when $n = 30$

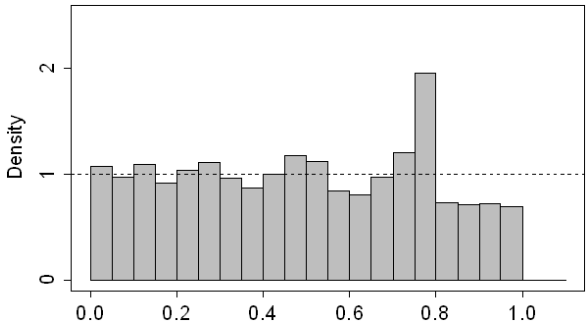

Figure 5: p-values (T'_{BC}) when $n = 100$

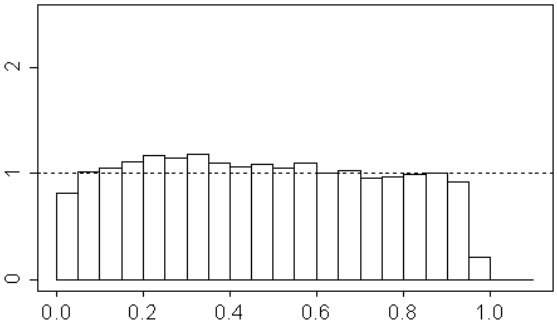

Figure 6: p-values ($\widehat{pv}(x)$) when $n = 100$

Figure 3 to 6: Histogram of 10,000 p-values from complete data (vertical axis: p-value, horizontal axis: density)

previously mentioned, we set four levels for sample size n and the percentage of censoring, respectively. Each simulation consists of 10,000 iterations.

In Figure 3 to 6, we provide four histograms of 10,000 p-values obtained under the situation where the percentage of censoring is 0% and the sample size $n = 30$ and 100. Figure 3 and 5 show histograms of 10,000 p-values based on T'_{BC}, and Figure 4 and 6 show ones based on $\widehat{pv}(x)$. We can consider that distributions of p-values obtained by T'_{BC} have larger variability than that of $\widehat{pv}(x)$. Especially, the density of T'_{BC} at around 0.8 is extremely large when $n = 30$. This reflects that, when $n = 30$ for complete data, two of the difference $|\widehat{S}_j^{\mathrm{lin}}(\widehat{M}_0) - 0.5|$, $j = 1.2$ are difficult to be zero simultaneously due to the composition of $\widehat{M}_0$ and $\widehat{S}_j^{\mathrm{lin}}(\widehat{M}_0)$. In other words, T'_{BC} does not still accomplish the continuity of asymptotic distribution. More important result in practice is that the proportion of p-values less than 0.05 for T'_{BC} is large while it is slightly small for $\widehat{pv}(x)$

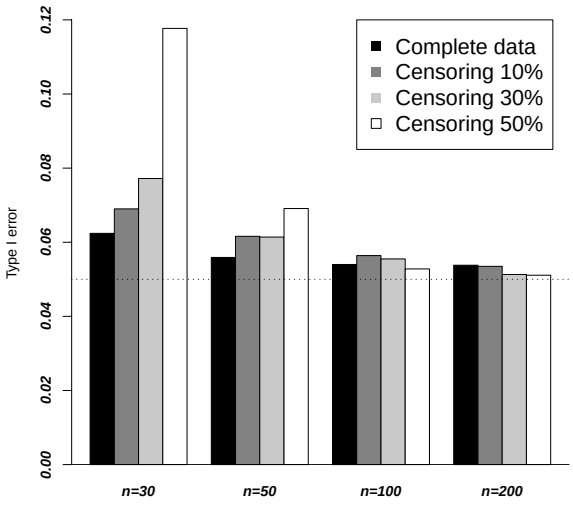

Figure 7: Empirical α of T'_{BC}

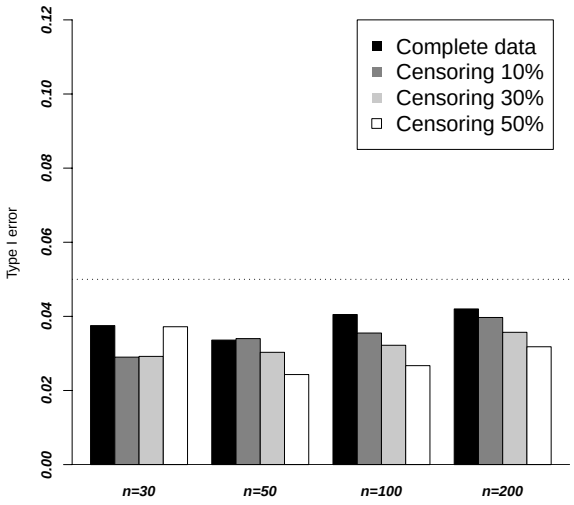

Figure 8: Empirical α of $\widehat{pv}(x)$

Figure 7 to 8: Empirical significance level based on 10,000 p-values for the nominal significance level of 0.05

on average. These biases are reduced in proportion to sample size n, for example, when $n = 100$ and 200, and their proportions closed to the nominal significance level of 0.05. We presented these results for complete data in Figure 3 to 6 in order to figure out the entire picture of the null distribution under the situation where no effect caused by censoring is existed. Other results based on the censored data also show the similar trends, namely, the distribution of p-values by T'_{BC} has larger variability than that of $\widehat{pv}(x)$, the proportion of p-values less than 0.05 for T'_{BC} is slightly large while it is slightly small for $\widehat{pv}(x)$. It is confirmed that the null distribution of $\widehat{pv}(x)$ is almost uniform except two-sided tails and it yields the conservativeness of testing, so we conclude that the proposed testing procedure and the associated calculation method has appropriate performance.

Next, we consider the proportion of p-values less than 0.05. We provide bar graphs of the empirical type I error (empirical α) against the nominal significance level of 0.05 which is defined as the number of p-values less than 0.05 of 10,000 p-values, for all combinations between sample size n and the proportion of censoring (16 cases in total). Figure 7 and 8 show testing results based on T'_{BC} and $\widehat{pv}(x)$ respectively (See also Table 1). Median test based on T'_{BC} shows the empirical α's are greater than 0.05 in all cases, and a tendency

Table 1: Empirical type I error

	Size 200		Size 100		Size 50		Size 30	
%Cens.	B&C	Proposed	B&C	Proposed	B&C	Proposed	B&C	Proposed
0%	0.0538	0.0420	0.0540	0.0405	0.0559	0.0336	0.0624	0.0375
10%	0.0535	0.0397	0.0564	0.0355	0.0616	0.0340	0.0690	0.0290
30%	0.0513	0.0357	0.0555	0.0322	0.0614	0.0303	0.0772	0.0292
50%	0.0511	0.0318	0.0528	0.0267	0.0691	0.0243	0.1177	0.0372

closing to the nominal significance level 0.05 with increasing n. And, another median test based on $\widehat{pv}(x)$ shows the empirical α's are less than 0.05 in all cases, and a tendency closing to 0.05 with increasing n. We notice that the empirical α of median test based on T'_{BC} exceeded 0.05 extremely as increasing the percentage of censoring in the small sample. Especially, a case of the percentage of censoring of 50% suggests a serious problem due to the empirical α beyond 0.1. For median test based on $\widehat{pv}(x)$, only a case of $n = 30$ shows a tendency that the empirical α is somewhat proportional to the percentage of censoring (with reasons considered in section 2.5.4). In other cases, the conservativeness of the test was enhanced by the way dealing with censoring which is proposed in section 2.5.3, so that the empirical α did not exceed the nominal significance level of 0.05.

3.1.2 Comparison of powers

According to the simulation model described previously, we compare the power of median tests based on T'_{BC} and $\widehat{pv}(x)$. For the alternative family of H_0, the difference of survival functions between sample 1 and 2 arise early in proportion to the value of a. And, the range for the difference in median survival times δ can be determined depending on that value. In this experiment, we set $a = 0.2$ so that δ can range from 0 to 100. Here, we also set $\max(\delta) = 90$ from a practical point view. The reason is that the survival function must drop vertically at a time-point with survival probability of $0.5 + a$ when $\delta = 100$. For the purpose of comparing the power and achieving the empirical significance level 0.05 for the test, we used the critical value obtained empirically by simulation with $\delta = 0$ as the rule of testing, based on the investigations for the null distribution in previous section.

We provide two power curves in cases of $n = 50$ and $n = 100$ in Figure 9 and 10, respectively (See also Table 2). In addition, Figure 9 and 10 show power curves for

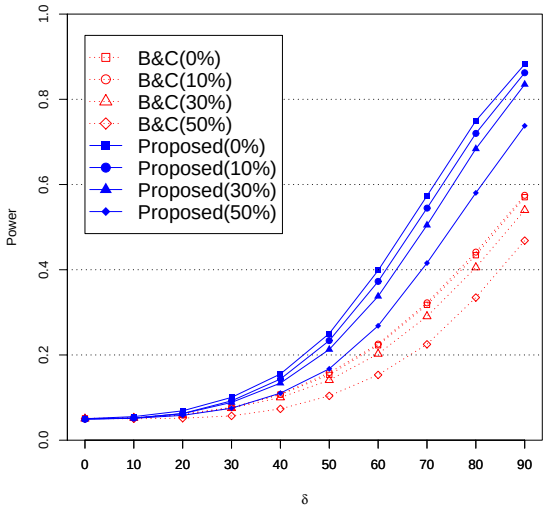

Figure 9: Power curve when $n = 50$

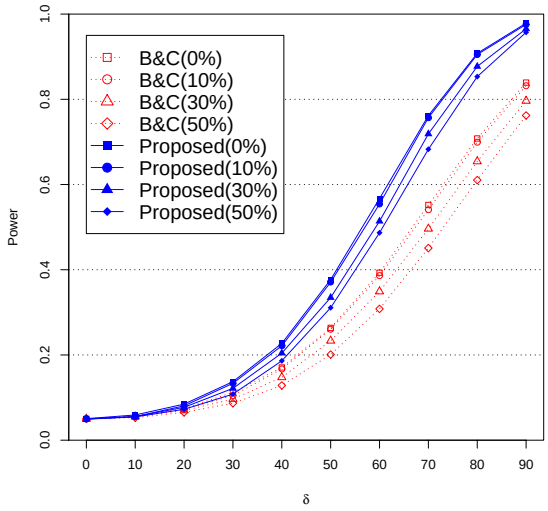

Figure 10: Power curve when $n = 100$

Figure 9-10: Power curves of T'_{BC}-based test (dashed) and $\widehat{pv}(x)$-based test (solid) for %censoring with 4 levels (horizontal axis: δ, vertical axis: Power)

T'_{BC}-based test (dashed line: B&C test) and $\widehat{pv}(x)$-based test (solid line: proposed test) simultaneously by the level of percentage of censoring (0%, 10%, 30% and 50%).

From Figure 9 and 10, we found that $\widehat{pv}(x)$-based test has higher power than T'_{BC}-based test consistently. With increase of the percentage of censoring, the power becomes smaller than that of complete data for each of testing procedures and sample sizes. However, the power of $\widehat{pv}(x)$-based test under 50% censoring is significantly higher than that of T'_{BC}-based test under no censoring. Furthermore, even if we use the nominal critical value 0.05 without such condition that the empirical α has to be 0.05, $\widehat{pv}(x)$-based test reached power of 0.8 with smaller δ compared to T'_{BC}-based test. As stated above, the proposed test is found to be superior to the test by Brookmeyer and Crowley (1982b) from point view of power (at least in this simulation model).

For the information, results for $n = 30$ are omitted since both median tests did not meet a power of 0.8 even complete data. Similar results were seen for $n = 200$ with ones for $n = 100$ while the difference in powers for both testing procedures was slightly reduced overall. (namely, $\widehat{pv}(x)$-based test under 50% censoring had higher power than T'_{BC}-based test under no censoring for all δ's greater than 40)

Table 2: Power comparisons

%Cens.	a	δ	Size 200		Size 100		Size 50		Size 30	
			B&C	Proposed	B&C	Proposed	B&C	Proposed	B&C	Proposed
0%	0.2	0	0.0496	0.0492	0.0502	0.0505	0.0513	0.0504	0.0498	0.0505
		10	0.0598	0.0570	0.0551	0.0592	0.0528	0.0553	0.0534	0.0518
		20	0.0961	0.1040	0.0742	0.0846	0.0616	0.0689	0.0591	0.0633
		30	0.1673	0.2020	0.1122	0.1374	0.0770	0.1004	0.0694	0.0858
		40	0.2852	0.3621	0.1707	0.2271	0.1072	0.1553	0.0884	0.1237
		50	0.4541	0.5786	0.2635	0.3769	0.1532	0.2503	0.1209	0.1941
		60	0.6514	0.7981	0.3930	0.5671	0.2227	0.3983	0.1650	0.2910
		70	0.8228	0.9338	0.5521	0.7617	0.3176	0.5734	0.2254	0.4202
		80	0.9342	0.9889	0.7081	0.9079	0.4341	0.7496	0.3085	0.5747
		90	0.9799	0.9985	0.8392	0.9786	0.5703	0.8824	0.4089	0.7086
10%	0.2	0	0.0502	0.0505	0.0509	0.0496	0.0505	0.0492	0.0502	0.0493
		10	0.0613	0.0631	0.0569	0.0554	0.0521	0.0519	0.0523	0.0500
		20	0.0974	0.1054	0.0724	0.0806	0.0588	0.0626	0.0572	0.0590
		30	0.1725	0.1965	0.1048	0.1335	0.0764	0.0923	0.0686	0.0768
		40	0.2839	0.3512	0.1672	0.2215	0.1071	0.1435	0.0890	0.1063
		50	0.4550	0.5654	0.2603	0.3709	0.1584	0.2337	0.1213	0.1713
		60	0.6443	0.7841	0.3859	0.5543	0.2256	0.3725	0.1668	0.2580
		70	0.8174	0.9301	0.5406	0.7568	0.3223	0.5444	0.2312	0.3695
		80	0.9299	0.9869	0.6995	0.9050	0.4413	0.7200	0.3175	0.5004
		90	0.9808	0.9990	0.8318	0.9757	0.5749	0.8623	0.4196	0.6043
30%	0.2	0	0.0510	0.0485	0.0501	0.0499	0.0503	0.0502	0.0497	0.0505
		10	0.0587	0.0583	0.0550	0.0554	0.0522	0.0517	0.0514	0.0529
		20	0.0874	0.0918	0.0697	0.0772	0.0590	0.0622	0.0565	0.0594
		30	0.1504	0.1725	0.0966	0.1216	0.0750	0.0886	0.0632	0.0730
		40	0.2470	0.3105	0.1475	0.2043	0.1002	0.1339	0.0772	0.0967
		50	0.3945	0.5062	0.2333	0.3347	0.1408	0.2129	0.1014	0.1398
		60	0.5779	0.7274	0.3489	0.5139	0.2025	0.3377	0.1385	0.1970
		70	0.7647	0.8999	0.4961	0.7188	0.2905	0.5046	0.1911	0.2653
		80	0.8990	0.9781	0.6542	0.8771	0.4053	0.6838	0.2644	0.3289
		90	0.9702	0.9980	0.7962	0.9654	0.5398	0.8350	0.3674	0.3520
50%	0.2	0	0.0498	0.0491	0.0502	0.0493	0.0505	0.0497	0.0510	0.0501
		10	0.0581	0.0590	0.0528	0.0546	0.0503	0.0513	0.0501	0.0494
		20	0.0807	0.0912	0.0653	0.0725	0.0516	0.0583	0.0470	0.0513
		30	0.1301	0.1614	0.0867	0.1081	0.0570	0.0750	0.0491	0.0556
		40	0.2152	0.2891	0.1284	0.1862	0.0735	0.1104	0.0518	0.0632
		50	0.3566	0.4762	0.2010	0.3109	0.1040	0.1675	0.0597	0.0780
		60	0.5287	0.6938	0.3082	0.4868	0.1532	0.2684	0.0779	0.0980
		70	0.7150	0.8772	0.4509	0.6828	0.2250	0.4157	0.1093	0.1215
		80	0.8693	0.9693	0.6104	0.8532	0.3348	0.5804	0.1615	0.1814
		90	0.9554	0.9970	0.7621	0.9572	0.4684	0.7378	0.2450	0.2678

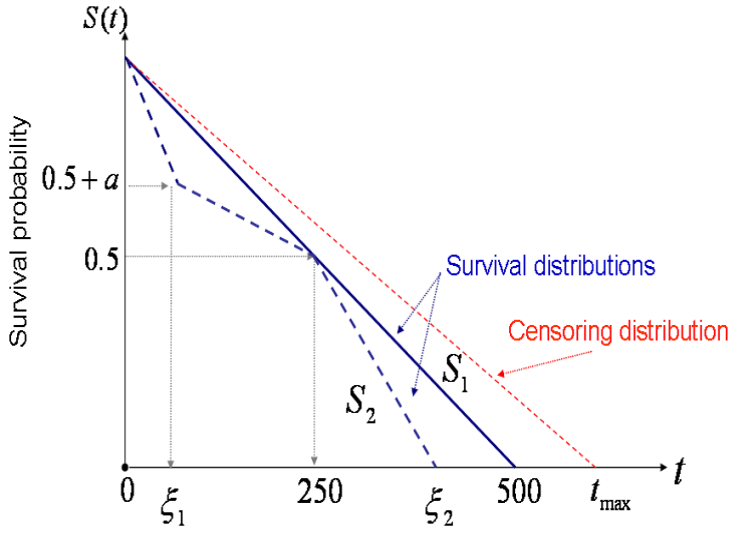

Figure 11: Graphical illustration of modified simulation

3.1.3 Power under a hypothesis of equal medians

One may be interested in powers of both testing procedures under a hypothesis $\{M_1^* = M_2^* \text{ and } S_1(t) \neq S_2(t), \exists t\}$. To achieve this, we modified simulation model slightly. Assume that distribution function $1 - S_1(t)$ is uniform on $[0, 500]$, and $S_2(t)$ is linear function with two branch time-points $(\xi_1, 0.5+a)$ and $(250, 0.5)$ (See Figure 11). Therefore, two survival distributions cross at the common median survival time $t = 250$. Thus, survival functions for each sample are written by

$$S_1(t) = 1 - \frac{t}{500}, \qquad S_2(t) = \begin{cases} 1 - (a - 0.5)t/100, & t \leq \xi_1 \\ c_1 t + c_2, & \xi_1 < t \leq 0.5 \\ c_3 t + c_4, & t > 0.5, \end{cases}$$

where $c_1 = -a/150$, $c_2 = 0.5 + 5a/3$, $c_3 = -0.5/(250 - d)$ and $c_4 = 0.5 - 250c_3$. Note that the slope of survival function after the common median in sample 2 is determined by the location of $\xi_2 = 500 - d$. In this simulation, we set $a = 0.25$, $\xi_1 = 100$ and $d \in \{0, 50, 100, 150, 200\}$ respectively. Also, we carry out the simulation for $n = 30$ and $n = 100$, respectively.

Power curves for each of sample size are provided in Figure 12 and 13. From the simulation results, we found that the power for each of median tests was increased as larger discrepancy of $S_1(t) \neq S_2(t)$ even if $M_1^* = M_2^*$. Of course, it is certain that

powers are rather low since this simulation was carried out under equal medians. As a characteristics of T'_{BC}-based test, the power was increased with increasing percentage of censoring. Similar trend has already been discussed in Figure 7 presenting bar graph of empirical type I error for the nominal significance level of 0.05 under H_0. And, it was suggested that the sign test would not be sensitive to the discrepancy of two survival distributions after median survival time. Meanwhile, as characteristics of $\widehat{pv}(x)$-based test, the power was decreased with increasing percentage of censoring. However, it was also suggested to have larger power with increased n. Similar trend is discussed in Figure 8. Comparing to the power with T'_{BC}-based test, although $\widehat{pv}(x)$-based test has smaller power than that of T'_{BC}-based test when $n = 30$, the former has larger power for large d when $n = 100$.

Table 3: Power comparisons under $\mathrm{H}_0^{\mathrm{m}}$

		Size 30		Size 100	
%Cens.	d	B&C	Proposed	B&C	Proposed
0%	0	0.0633	0.0395	0.0554	0.0428
	50	0.0640	0.0417	0.0548	0.0459
	100	0.0665	0.0532	0.0560	0.0560
	150	0.0710	0.0767	0.0631	0.0760
	200	0.0825	0.1241	0.0774	0.1269
10%	0	0.0700	0.0319	0.0573	0.0393
	50	0.0716	0.0405	0.0583	0.0438
	100	0.0747	0.0508	0.0627	0.0526
	150	0.0810	0.0736	0.0682	0.0732
	200	0.0936	0.1240	0.0816	0.1193
30%	0	0.0783	0.0300	0.0532	0.0363
	50	0.0781	0.0347	0.0550	0.0401
	100	0.0786	0.0473	0.0582	0.0483
	150	0.0839	0.0593	0.0650	0.0709
	200	0.0963	0.0804	0.0759	0.1180
50%	0	0.1175	0.0379	0.0535	0.0285
	50	0.1141	0.0380	0.0542	0.0329
	100	0.1122	0.0417	0.0570	0.0428
	150	0.1085	0.0515	0.0615	0.0650
	200	0.1091	0.0679	0.0722	0.1123

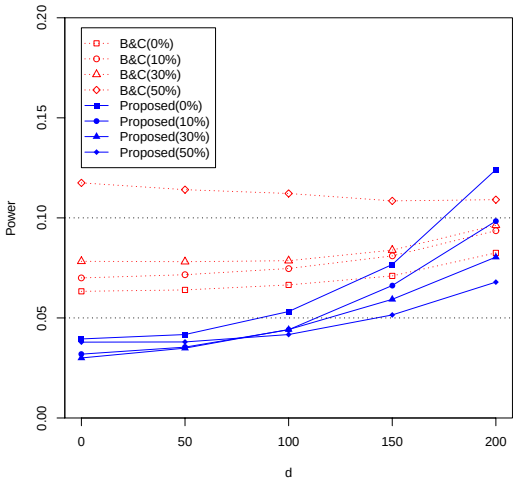

Figure 12: Power curve when $n = 30$

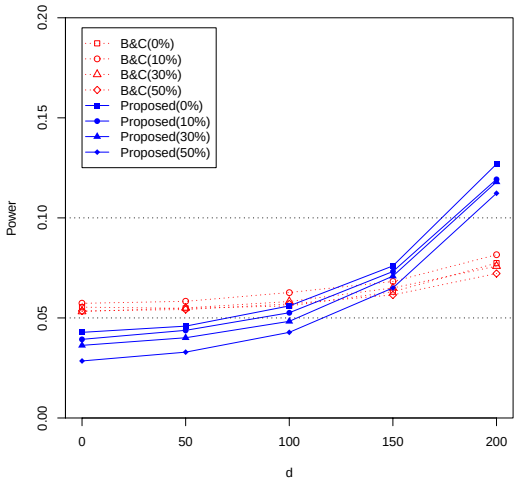

Figure 13: Power curve when $n = 100$

Figure 12-13: Power curves of T'_{BC}-based test (dashed) and $\widehat{pv}(x)$-based test (solid) for %censoring with 4 levels (horizontal axis: δ, vertical axis: Power)

3.2 Simulation study 2

In section 3.1, a simulation study was carried out to investigate the performances for T'_{BC}-based test and $\widehat{pv}(x)$-based test. In this section, we carry out comprehensive simulation study to investigate the performances with respect to the null distribution, empirical type I error and power for each of four median tests introduced in Section 2. In order to carry out this simulation study, we utilized several underlying survival distributions given in Table 4.

In the simulation, four levels of the percentage of censoring, namely, 0%, 10%, 30% and 50% are considered respectively. The percentage of censoring is adjusted by the value of scale parameter in assumed exponential distribution as the censoring distribution. We

Table 4: Assumed underlying survival distributions in the simulation

Survival distribution	p.d.f.
Uniform	$f(t) = \frac{1}{b-a} \quad (a \le t \le b)$
Exponential(λ)	$f(t;\lambda) = \lambda \exp(-\lambda t)$
Weibull(λ,ρ)	$f(t;\lambda,\rho) = \lambda\rho(\lambda t)^{\rho-1} \exp\{-(\lambda t)^{\rho}\}$
Gamma(κ, λ)	$f(t;\kappa,\lambda) = \frac{\lambda(\lambda t)^{\kappa-1}\exp(-\lambda t)}{\Gamma(\kappa)}$
Log-normal(μ, σ^2)	$f(t;\mu,\sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp\{-\frac{1}{2\sigma^2}(\log t - \mu)^2\}\frac{1}{t}$

set equal sample size for each of two samples, that is, $(n_1 = n_2 =)15$ and 50 per sample are considered respectively.

With respect to the examination of the null distribution and empirical type I error for each of median tests, only Exponential($\lambda = 1$) and Log-normal($\mu = 0, \sigma = 1$) were used. With respect to the examination of the power, the following shift alternative was considered, namely,

$$\mathrm{H}_1 : S_1(t) = S_2(t + \delta).$$

As a result, δ was adjusted for each of underlying survival distributions so that at least one testing procedure could have adequate power when $n = 100$. That is, the same δ was used within the same underlying survival distribution regardless of the percentage of censoring and sample size. For all simulation we utilized 10,000 iterations except bootstrap median test. For bootstrap median test only, we utilized 1,000 iterations with $B = 1,000$ per iteration.

3.2.1 Null distribution for each of median tests

In order to investigate the null distributions of T'_{BC}, T_{ELR}, $\widehat{p}_{boot}$ and $\widehat{pv}(x)$, survival data was generated from Exponential(1) and Log-normal(0,1) under H_0.

We provide four histograms of p-values obtained under the situation (only for Exponential(1)) where the percentage of censoring is 0% and the sample size $n = 30$ and 100 in Figure 14 to 21. Histograms of 10,000 p-values based on T'_{BC} are given in Figure 14 and 18, ones based on T_{ELR} in Figure 15 and 19, ones based on $\widehat{p}_{boot}$ in Figure 16 and 20, and ones based on $\widehat{pv}(x)$ in Figure 17 and 21. With respect to null distributions of T'_{BC}, they show the similar trends as discussed in Simulation 1. That is, the density of T'_{BC} at around 0.8 is large when $n = 30$. As discussed in Simulation 1, we consider that T'_{BC} does not still accomplish the continuity of asymptotic distribution. Also, the proportion of p-values less than 0.05 for T'_{BC} is large compared to $\widehat{p}_{boot}$ and $\widehat{pv}(x)$ on average. With respect to null distributions of T_{ELR}, it is suggested that the null distribution of T_{ELR} is not continuous, especially when $n = 30$. It is also suggested that null distribution of T_{ELR} tends to be continuous as the sample size increased. The reason why the density of T_{ELR} at around 1.0 is relatively large, is that T_{ELR} provides p-value of 1.0 when two observed

median survival times are equal by chance. Since the ratio between unconstrained and constrained maximum log likelihoods is zero, therefore $T_{\mathrm{ELR}} = 0$. For $\widehat{p}_{boot}$ and $\widehat{pv}(x)$, the shape of null distributions are very similar. That is, their null distributions are almost uniform except both tails.

Based on these discussions, it is confirmed that the null distributions of $\widehat{p}_{boot}$ and $\widehat{pv}(x)$ yields the conservativeness of testing. On the other hand, T_{ELR} and T'_{BC} would not yield the conservativeness, especially for small samples. In a sense, T_{ELR} may yield the conservativeness of testing, however there is room for further research into the shape of its null distribution.

Next, we discuss the proportion of p-values less than 0.05. We provide bar graphs of the empirical type I error (empirical α) against the nominal significance level of 0.05 which is defined as the number of p-values less than 0.05 of 10,000 p-values, for all combinations between sample size n and the percentage of censoring. Figure 22 to 25 shows the empirical type I error for each of median tests (See also Table 5). The generalized sign test based on T'_{BC} shows the empirical α's are greater than 0.05 in all cases, and a tendency closing to the nominal significance level 0.05 with increasing n. Also, its empirical α exceeded 0.05 as increasing the percentage of censoring for small samples. The other median tests based on T_{ELR} and $\widehat{pv}(x)$ show the empirical α's are less than 0.05 in all cases, and a tendency closing to 0.05 with increasing n when $n = 100$. Bootstrap median test based on $\widehat{p}_{boot}$ shows the empirical α's are less than 0.05 in all cases too, however, no consistent tendency was seen.

3.2.2 Power for each of median tests

We utilized the standardized distributions for most of assumed underlying survival distributions, namely, Uniform($a = 0, b = 1$), Exponential($\lambda = 1$), Log-Normal($\mu = 0, \sigma = 1$) (log-transform of standard normal). In addition, Weibull($\lambda = 1, \rho = 2$) and Gamma($\kappa = 2, \lambda = 1$) were utilized.

Table 6 shows a simple comparison under shifted alternative of δ. The value of δ was adjusted for each of underlying survival distributions so that at least one median test could have adequate power when $n = 100$. For $n = 100$, each of median tests have almost the same power except the empirical likelihood ratio test. The empirical likelihood ratio

test has slightly smaller power compared to other median tests. We consider that this was caused by its conservativeness as seen in the investigation of the empirical type I error. The same tendency was also shown when $n = 30$. For $n = 30$, the generalized sign test has higher power than other median tests. Considerable reasons are: vertical difference $|\widehat{S}_j^{\mathrm{lin}}(\widehat{M}_0) - 0.5|$ is sensitive compared to the horizontal differences in medians defined by several ways for small samples, and the characteristic of simulation design, namely, shift model. Through the investigation of powers under shift model, bootstrap median test and proposed median test tend to be powerful as increasing n.

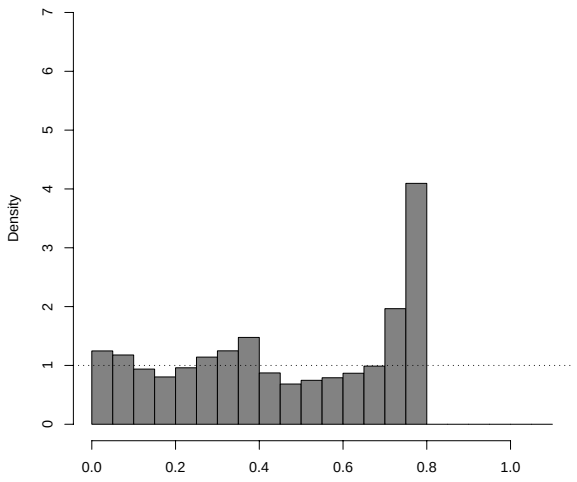

Figure 14: p-values (T'_{BC}) when $n = 30$

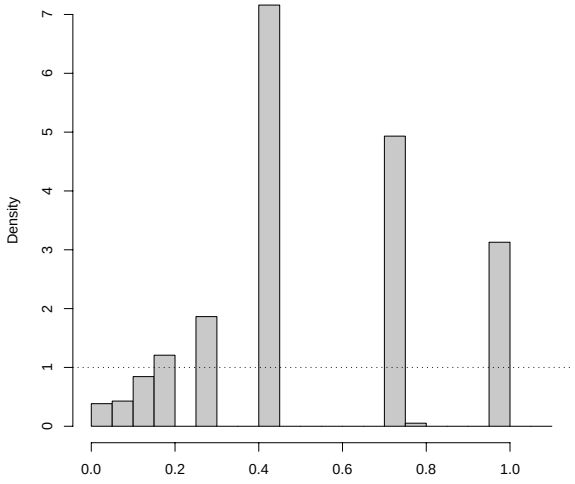

Figure 15: p-values (T_{ELR}) when $n = 30$

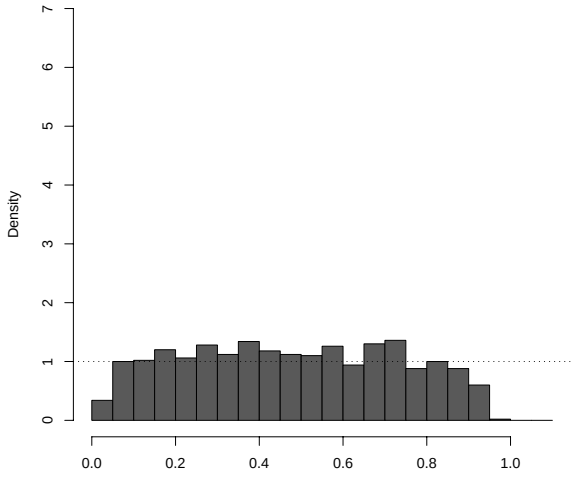

Figure 16: p-values ($\widehat{p}_{boot}$) when $n = 30$

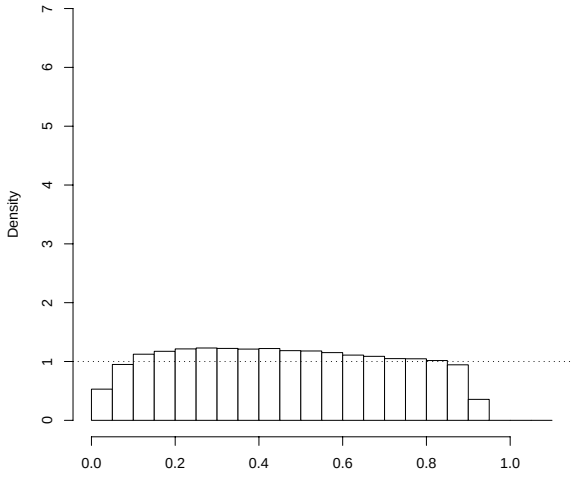

Figure 17: p-values ($\widehat{pv}(x)$) when $n = 30$

Figure 14 to 17: Histogram of p-values from complete data when $n = 30$
(vertical axis: density, horizontal axis: p-value)

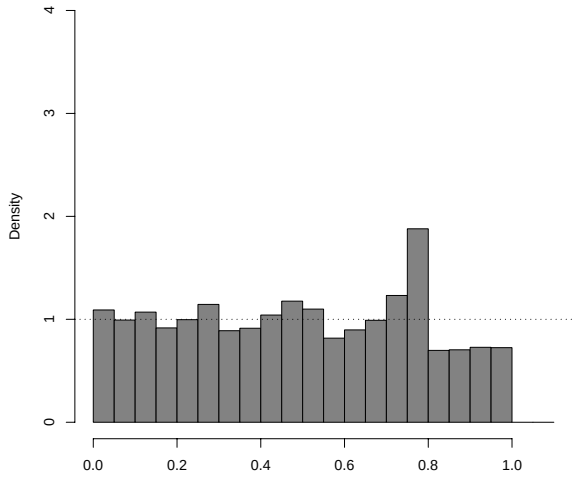

Figure 18: p-values (T'_{BC}) when $n = 100$

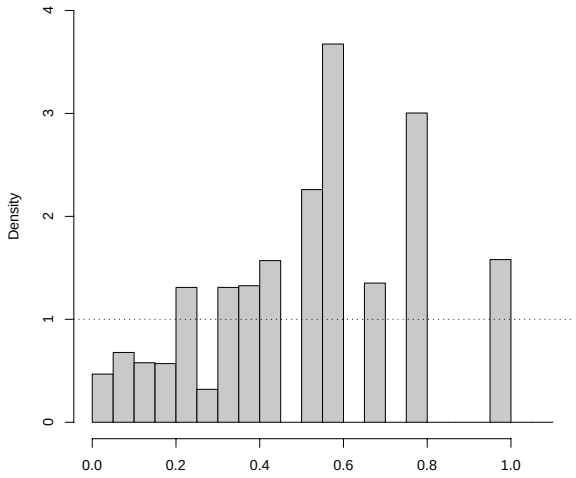

Figure 19: p-values (T_{ELR}) when $n = 100$

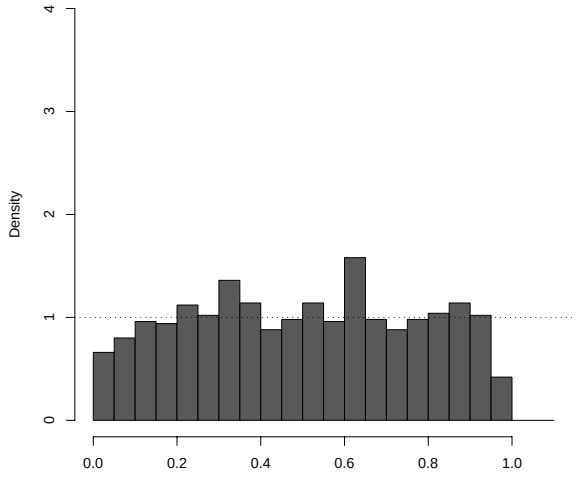

Figure 20: p-values ($\widehat{p}_{boot}$) when $n = 100$

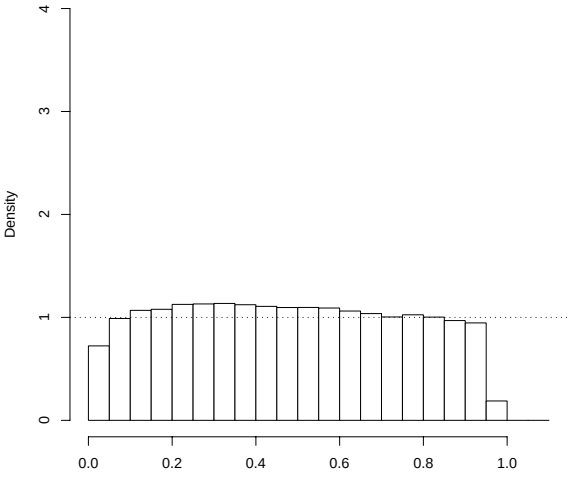

Figure 21: p-values ($\widehat{pv}(x)$) when $n = 100$

Figure 18 to 21: Histogram of p-values from complete data when $n = 100$
(vertical axis: density, horizontal axis: p-value)

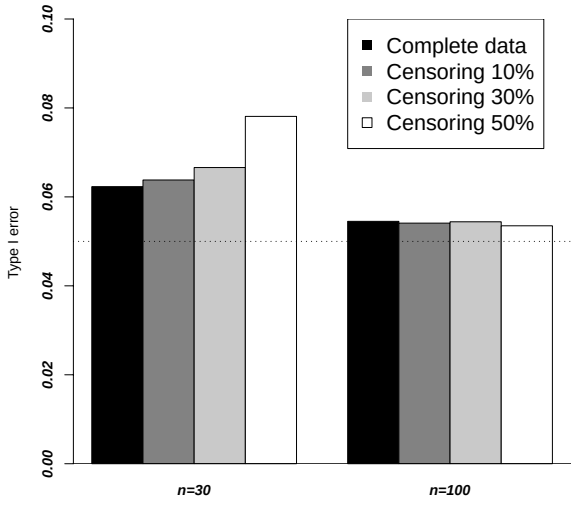

Figure 22: Empirical α of T'_{BC}

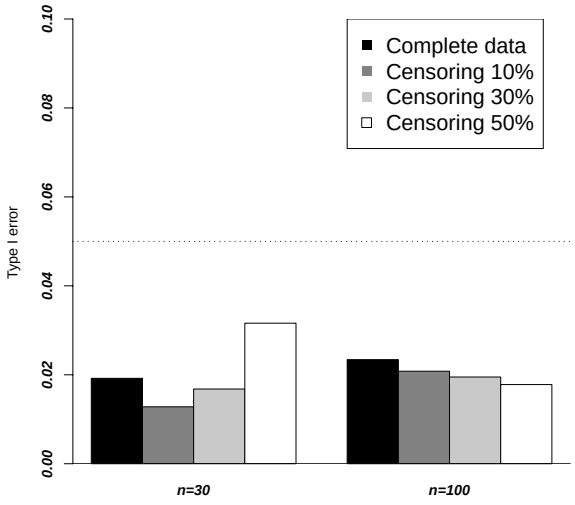

Figure 23: Empirical α of T_{ELR}

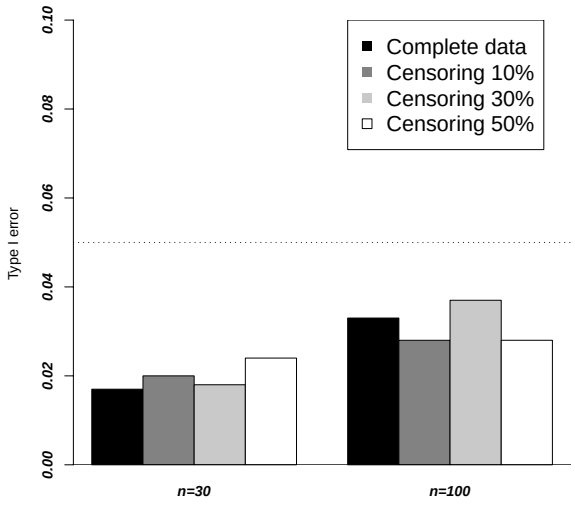

Figure 24: Empirical α of $\widehat{p}_{boot}$

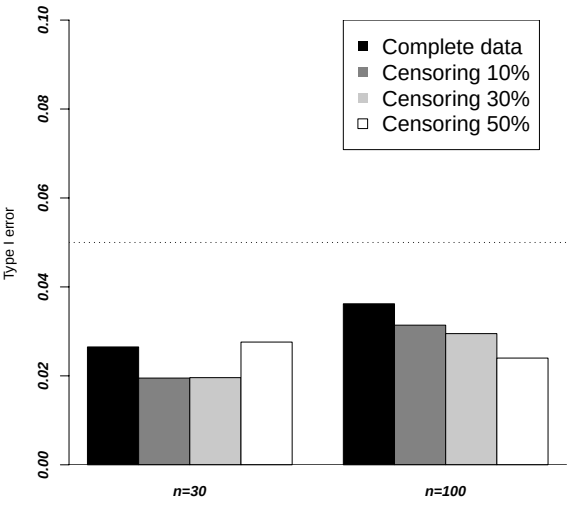

Figure 25: Empirical α of $\widehat{pv}(x)$

Figure 22 to 25: Empirical significance level based on 10,000 p-values for the nominal significance level of 0.05

Table 5: Empirical type I error for each of median tests

Sample size	%Cens.	Type of Median test	Exponential(1)	Log-normal(0,1)
30	0%	B&C	0.0623	0.0624
		Empirical likelihood ratio test	0.0192	0.0171
		Bootstrap median test	0.0170	0.0240
		Proposed median test	0.0265	0.0234
	10%	B&C	0.0638	0.0646
		Empirical likelihood ratio test	0.0128	0.0137
		Bootstrap median test	0.0200	0.0200
		Proposed median test	0.0195	0.0165
	30%	B&C	0.0666	0.0668
		Empirical likelihood ratio test	0.0168	0.0175
		Bootstrap median test	0.0180	0.0160
		Proposed median test	0.0196	0.0170
	50%	B&C	0.0781	0.0782
		Empirical likelihood ratio test	0.0316	0.0298
		Bootstrap median test	0.0240	0.0390
		Proposed median test	0.0276	0.0261
100	0%	B&C	0.0545	0.0547
		Empirical likelihood ratio test	0.0234	0.0232
		Bootstrap median test	0.0330	0.0350
		Proposed median test	0.0362	0.0338
	10%	B&C	0.0541	0.0538
		Empirical likelihood ratio test	0.0208	0.0200
		Bootstrap median test	0.0280	0.0350
		Proposed median test	0.0314	0.0292
	30%	B&C	0.0544	0.0540
		Empirical likelihood ratio test	0.0195	0.0187
		Bootstrap median test	0.0370	0.0260
		Proposed median test	0.0295	0.0275
	50%	B&C	0.0535	0.0546
		Empirical likelihood ratio test	0.0178	0.0161
		Bootstrap median test	0.0280	0.0200
		Proposed median test	0.0240	0.0222

Table 6: Powers under shift alternative

Sample size	%Cens.	Type of median test	Underlying survival distributions: Uni $\delta = 0.5$	Exp(1) $\delta = 0.5$	Wei(1,2) $\delta = 0.4$	Gam(2,1) $\delta = 1$	LN(0,1) $\delta = 0.7$
30	0%	B&C	0.4723	0.4179	0.5090	0.4777	0.4127
		ELR	0.3783	0.2513	0.3804	0.3067	0.2148
		Bootstrap	0.6020	0.3000	0.4200	0.3680	0.2830
		Proposed	0.6306	0.3207	0.4364	0.3870	0.2692
	10%	B&C	0.4815	0.4106	0.5062	0.4636	0.4130
		ELR	0.3524	0.2407	0.2987	0.2874	0.1884
		Bootstrap	0.5610	0.2790	0.3700	0.3330	0.2220
		Proposed	0.5466	0.2815	0.3872	0.3263	0.2034
	30%	B&C	0.4411	0.3906	0.4629	0.4547	0.3875
		ELR	0.3112	0.1965	0.2436	0.2491	0.1325
		Bootstrap	0.4350	0.2410	0.2950	0.3150	0.1770
		Proposed	0.4078	0.2428	0.3261	0.3146	0.1529
	50%	B&C	0.3986	0.3603	0.4100	0.4065	0.3654
		ELR	0.2857	0.1669	0.2055	0.1943	0.1183
		Bootstrap	0.3500	0.1900	0.2450	0.2410	0.1890
		Proposed	0.3255	0.2005	0.2556	0.2564	0.1418
100	0%	B&C	0.8795	0.7956	0.9230	0.8848	0.8555
		ELR	0.7886	0.7215	0.8545	0.7796	0.8064
		Bootstrap	0.9050	0.7960	0.9310	0.8980	0.8650
		Proposed	0.9141	0.8160	0.9354	0.9023	0.8533
	10%	B&C	0.8529	0.7870	0.9128	0.8613	0.8385
		ELR	0.7492	0.7032	0.8279	0.7263	0.7528
		Bootstrap	0.8770	0.7740	0.9070	0.8703	0.8220
		Proposed	0.8791	0.7935	0.9183	0.8663	0.8056
	30%	B&C	0.8007	0.7532	0.8685	0.8532	0.7964
		ELR	0.6608	0.6687	0.7513	0.7169	0.6733
		Bootstrap	0.8070	0.7520	0.8740	0.8495	0.7660
		Proposed	0.8311	0.7536	0.8711	0.8551	0.7457
	50%	B&C	0.7211	0.6716	0.8024	0.7972	0.7338
		ELR	0.5466	0.5866	0.6375	0.6187	0.5769
		Bootstrap	0.7380	0.6450	0.7750	0.7924	0.6580
		Proposed	0.7620	0.6460	0.7922	0.7895	0.6294

4. Case Studies and Numerical Investigation

In this chapter, we carried out two case studies to investigate the behaviors for each of four median tests, that is, generalized sign test, empirical likelihood ratio test, bootstrap median test and proposed median test. As references, log-rank test (Cox-Mantel) and Generalized Wilcoxon test (Modified Peto-Peto) were also applied as well. The reasons why such rank tests were involved are to investigate how sensitive are such rank tests for the difference of survival distributions compared to median tests and to give a caution to interpret obtained result in particular case, such as crossing survival distributions.

First case was selected with the reason that survival data provides the standard two survival distributions with no cross between them. Second case was selected with the reason that two survival distributions cross at a time-point (beyond two medians).

4.1 Case 1: Survival data of patients with tongue cancer

Survival data was collected to investigate the effects of ploidy (aneuploid or diploid) on the prognosis (survival time) of patients with cancers of the tongue (Sickle-Santanello *et al*, 1988). Let sample 1 and 2 be the aneuploid tumors group with 52 patients and the diploid tumors group with 28 patients, respectively. The number of censored observations for each sample was 21 and 6 respectively.

Table 7 shows the results based on both testing procedures as well as ones based on the log-rank test and the generalized Wilcoxon test as reference. The generalized Wilcoxon test provided a smaller p-value than that of the log-rank test. This is because relatively-large differences between two survival curves are seen for small t.

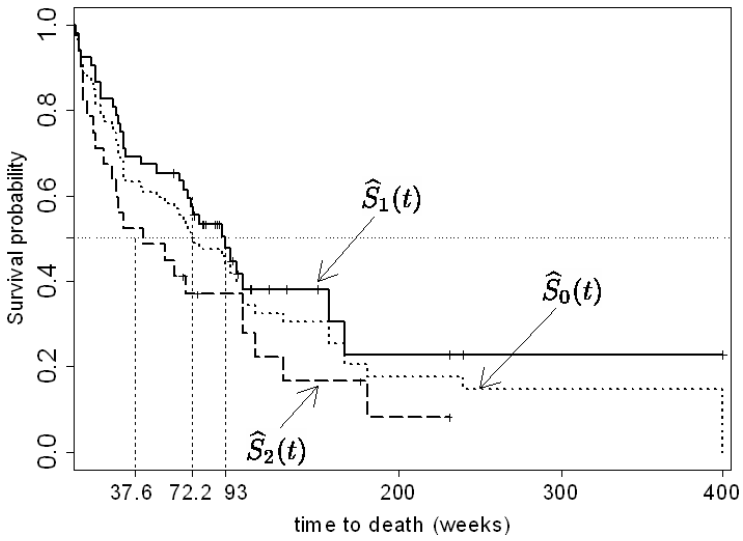

Figure 26: Kaplan-Meier estimates $\widehat{S}_1(t)$ (Aneuploid Tumors: solid line) and $\widehat{S}_2(t)$ (Diploid Tumors: dashed line), and weighted Kaplan-Meier estimate $\widehat{S}_0(t)$ (dotted line)

Let $\widehat{M_0}$ be the estimated median survival time under H_0 for which the linear interpolation of $\widehat{S}_0(t)$ meets a survival probability of 0.5. The generalized sign test evaluates the difference between an estimate of survival probability $\widehat{S}_j^{\rm lin}(\widehat{M_0})$ ($j = 1$ or 2) based on the linear interpolation and 0.5. For this data, it was estimated that $\widehat{M_0} = 72.2$ (in week), $\widehat{S}_1(\widehat{M_0})$ =0.569 and $\widehat{S}_2(\widehat{M_0})$ =0.362. Since $|\widehat{S}_2^{\rm lin}(\widehat{M_0}) - 0.5|$ is greater than $|\widehat{S}_1^{\rm lin}(\widehat{M_0}) - 0.5|$, the p-value from the test statistic $T_{\rm BC}$ based on $\widehat{S}_1^{\rm lin}(\widehat{M_0})$ is greater than one based on $\widehat{S}_2^{\rm lin}(\widehat{M_0})$ (p-value is 0.0816 for sample 1 while 0.0634 for sample 2). In practice, an issue is which p-value should be used. To maintain their balance, we used the unified test statistic $T'_{\rm BC}$ which provides a p-value of 0.0692.

Table 7: p-values for each of testing procedures in tongue cancer data

Type of test:	Procedure	χ^2	p-value
Median test:	Generalized sign test		
	Test statistic $T_{\rm BC}$ based on sample 1	3.032	0.0816
	Test statistic $T_{\rm BC}$ based on sample 2	3.447	0.0634
	Unified test statistic $T'_{\rm BC}$	3.302	0.0692
Median test:	Empirical likelihood ratio test	2.048	0.1523
Median test:	Bootstrap median test	-	0.0900
Median test:	Proposed test	-	0.0722
Log-rank test:	Cox-Mantel	2.790	0.0949
Generalized Wilcoxon test:	Modified Peto-Peto	3.306	0.0690

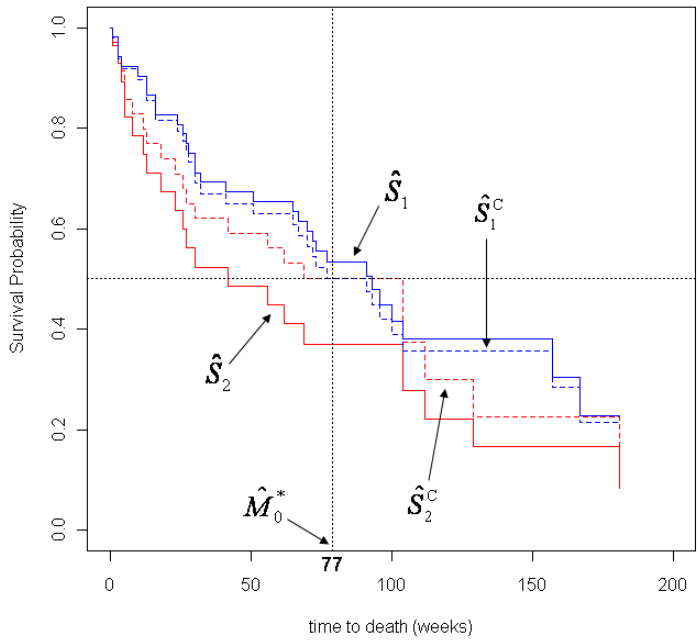

Figure 27: Kaplan-Meier estimates $\widehat{S}_j(t)$ and the constraint Kaplan-Meier estimates $\widehat{S}_j^c(t)$, $j = 1, 2$ (dotted line) for tongue cancer data

For the empirical likelihood ratio test, the unconstrained maximum log likelihood value was $\log L_u(\widehat{\lambda}_j(t)) = -205.0573$. In order to estimate the common median under H_0^m, the estimates of median defined by $\inf\{t : \widehat{S}_j(t) < 0.5\}$ $(j = 1, 2)$ were used. Each of estimated medians was $\widehat{M}_1 = 42$ and $\widehat{M}_2 = 93$, respectively. Among death times between 42 and 93, the constrained maximum log likelihood was obtained at 77, so the common median under H_0^m was estimated as $\widehat{M}_0^* = 77$. Here, through the Newton-Raphson method, the estimates of Lagrangian parameters at $\widehat{M}_0^*$ were $\widehat{\alpha}_1 = 6.91$ and $\widehat{\alpha}_2 = 3.50$ associated with $\phi_1 = 1$ and $\phi_2 = -1$, respectively. As a result, the constrained maximum log likelihood value was $\log L_c(\widehat{\lambda}_j^*(t)) = -206.0816$, and the empirical likelihood ratio test statistic T_{ELR} yields a p-value of 0.1523. Obtained p-value was the larger than that of any other median tests. One of the reasons for this is that this test procedure is considered more conservative compared to other median tests through simulation study 2. The behavior of the constrained log likelihoods between $\widehat{M}_1$ and $\widehat{M}_2$ is illustrated in Figure 1. In addition, the constrained Kaplan-Meier estimates with common median $\widehat{M}_0^* = 77$ are given in Figure 27.

The bootstrap median test was applied with $B = 1,000$. For this testing procedure, observed difference in medians to be used for the testing was $93 - 42 = 51$. Out of 1,000 bootstrap samples, the number of bootstrap samples providing a difference in medians greater than or equal to 51 were 90. Therefore, $\widehat{p}_{boot}$ was estimated as 0.0900.

For the proposed median test, estimated median survival times $\widehat{M}_j$ $(j = 1, 2)$ defined by (2.16) are 93 and 37.6, respectively. Based on these estimates, observed difference in median survival times was 55.4 which yields a p-value of 0.0722, namely, $\widehat{pv}(55.4) = 0.0722$. Adjusted sample sizes due to the censoring were $(n_1', n_2') \approx (27.1, 50.3)$. The value of $\widehat{pv}(55.4)$ is slightly greater than that of T'_{BC}. One of the reasons for this is that the testing procedure based on $\widehat{pv}(x)$ with no adjustment of α is conservative via investigations of the null distributions based on the simulation study. Further discussions on the Monte-Carlo method to adjust the critical value so that the empirical α can become 0.05 based on actual survival data are problems to be resolved in the future.

4.2 Case 2: Survival data of patients with gastric cancer

Survival data was observed for patients with locally unresectable gastric cancer in a randomized clinical trial (Stablein and Koutrouvelis, 1985). Either of two treatments (chemotherapy or chemotherapy plus radiotherapy) was assigned to them randomly. Let sample 1 and 2 be the chemotherapy group with 45 patients and the combination therapy (chemotherapy plus radiotherapy) group with 45 patients, respectively. The number of censored observations for each sample was 2 and 6 respectively.

The Kaplan-Meier estimates $\widehat{S}_1(t)$ (solid line) and $\widehat{S}_2(t)$ (dashed line) for each sample, and a weighted Kaplan-Meier estimate for them are provided in Figure 28. Table 8 shows the testing results similar to Table 7 for the first case. From Figure 28, we can see that the differences between $\widehat{S}_1(t)$ and $\widehat{S}_2(t)$ are large relatively in the beginning (in favor of $\widehat{S}_1(t)$), then both functions cross at a time-point and the advantage was reversed finally (in favor of $\widehat{S}_2(t)$). Based on such data, the generalized Wilcoxon test gives a smaller p-value while a p-value by log-rank test is considerably large, as indicated in Table 8.

As required information for the generalized sign test, M_0 is estimated as $\widehat{M}_0 = 401$

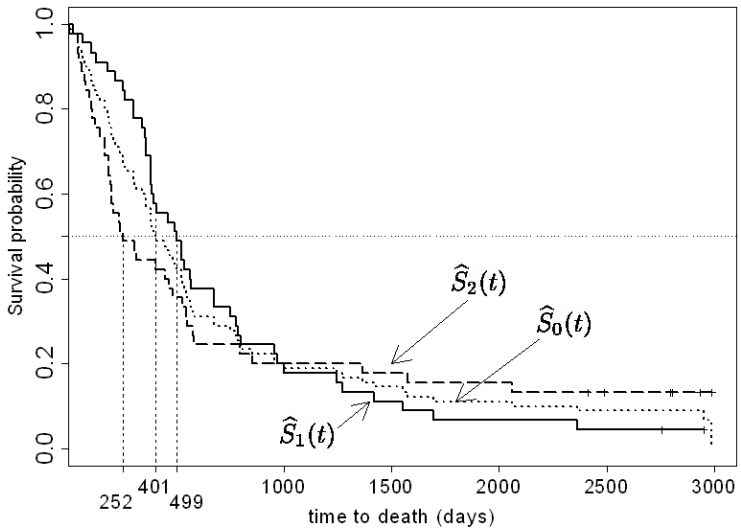

Figure 28: Kaplan-Meier estimates $\widehat{S}_1(t)$ (Chemotherapy only: solid line) and $\widehat{S}_2(t)$ (Combination therapy: dashed line), and weighted Kaplan-Meier estimate $\widehat{S}_0(t)$ (dotted line)

in days, and the estimates based on the linear interpolation of survival probability are $\widehat{S}_1^{\rm lin}(\widehat{M}_0)$ =0.566 and $\widehat{S}_2^{\rm lin}(\widehat{M}_0)$ =0.422, respectively. Since $\widehat{S}_2^{\rm lin}(\widehat{M}_0)$ has a larger difference with 0.5 compared to $\widehat{S}_1^{\rm lin}(\widehat{M}_0)$, test statistic based on $T_{\rm BC}$ for sample 2 is greater than that for sample 1. The unified test statistic $T'_{\rm BC}$ provides a p-value of 0.1624.

For the empirical likelihood ratio test, the unconstrained maximum log likelihood value was $\log L_{\rm u}(\widehat{\lambda}_j(t)) = -327.6901$. And, the estimates of median defined by $\inf\{t : \widehat{S}_j(t) < 0.5\}$ were $\widehat{M}_1 = 489$ and $\widehat{M}_2 = 254$. Among death times between 254 and 489, the constrained maximum log likelihood was obtained at 401, so the common median under $\mathrm{H}_0^{\rm m}$ was estimated as $\widehat{M}_0^* = 401$. Here, the estimates of Lagrangian parameters at $\widehat{M}_0^*$

Table 8: p-values for each of testing procedures in gastric cancer data

Type of test:	Procedure	χ^2	p-value
Median test:	Generalized sign test		
	Test statistic $T_{\rm BC}$ based on sample 1	1.653	0.1985
	Test statistic $T_{\rm BC}$ based on sample 2	2.250	0.1336
	Unified test statistic $T'_{\rm BC}$	1.952	0.1624
Median test:	Empirical likelihood ratio test	1.650	0.1990
Median test:	Bootstrap median test	-	0.1340
Median test:	Proposed test	-	0.0138
Log-rank test:	Cox-Mantel	0.232	0.6301
Generalized Wilcoxon test:	Modified Peto-Peto	3.997	0.0456

were $\widehat{\alpha}_1 = 6.91$ and $\widehat{\alpha}_2 = 3.50$ associated with $\phi_1 = 1$ and $\phi_2 = -1$, respectively. As a result, the constrained maximum log likelihood value was $\log L_{\rm c}(\widehat{\lambda}_j^*(t)) = -328.5151$, and the empirical likelihood ratio test statistic $T_{\rm ELR}$ yields a p-value of 0.1990. We note that $\widehat{M}_0^*$ under $\mathrm{H}_0^{\rm m}$ and $\widehat{M}_0$ under H_0 in the generalized sign test were equal. In this situation, the generalized sign test provided smaller p-value than that of the empirical likelihood ratio test. The behavior of the constrained log likelihoods between $\widehat{M}_1$ and $\widehat{M}_2$ is illustrated in Figure 29. In addition, the constrained Kaplan-Meier estimates with common median $\widehat{M}_0^* = 401$ are given in Figure 30.

For the bootstrap median test, observed difference in medians to be evaluated was $489 - 254 = 235$ according to estimated medians in the empirical likelihood ratio test. Out of 1,000 bootstrap samples, the number of bootstrap samples providing a difference in medians greater than or equal to 235 was 134. Therefore, $\widehat{p}_{boot} = 0.1340$.

For the proposed median test, meanwhile, the estimates of the median survival time for each group $\widehat{M}_1$ and $\widehat{M}_2$ defined by (2.16) are 499 and 252 (days), respectively. Because no survival time is censored prior to $\widehat{M}_j$, $j = 1, 2$ for both samples, we can define that $(n_1', n_2') = (n_1, n_2)$. Using these information, p-value by the proposed testing procedure was calculated as $\widehat{pv}(247) = 0.0138$ based on the observed difference in median survival times $\widehat{M}_1 - \widehat{M}_2 = 247$ (day).

Proposed median test achieved significance level of 0.05 while the generalized sign test did not. One of the reasons is that proposed median test would have higher power compared to other median tests under large sample and few censored observations as indicated in the simulation study 2. Another reason would be that the vertical difference $|\widehat{S}_j^{\rm lin}(\widehat{M}_0) - 0.5|$ on the survival curve is underestimated compared to the horizontal difference $|\widehat{M}_1 - \widehat{M}_2|$ on the scale of detecting the difference, due to the character of this data that the survival probabilities change rapidly around the estimated survival times.

Whether or not it be true, the important thing in the application is that how do we choose the measure for the comparison in survival distributions. And, note that we can not interpret the relative merits for two survival distributions by just looking at the size of p-value. The result by the generalized wilcoxon test meets the significance level of 0.05, but it can be hardly said that the result is easy to interpret under this case. On the other hand, the interpretation of results based on the median test is clear. If

one's most important and interesting objective is to compare median survival times, the proposed testing procedure has a great deal of potential to be applied since the proposed testing procedure detected the statistical significant difference in median survival times even though other median tests could not detect it in this case.

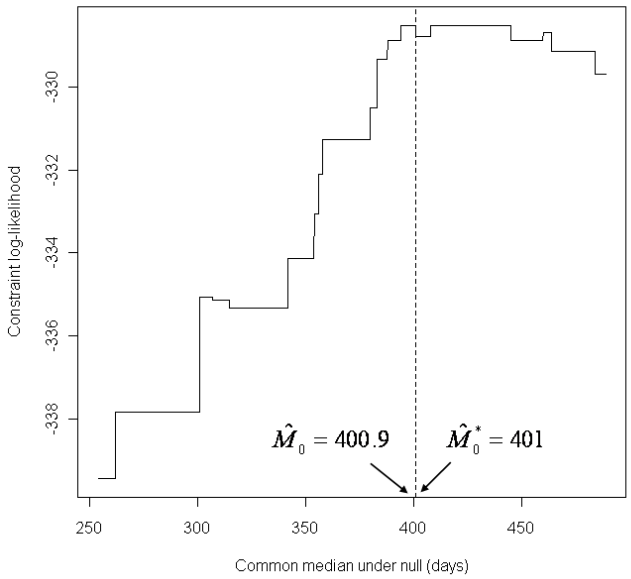

Figure 29: The behavior of the constrained log likelihoods for gastric cancer data

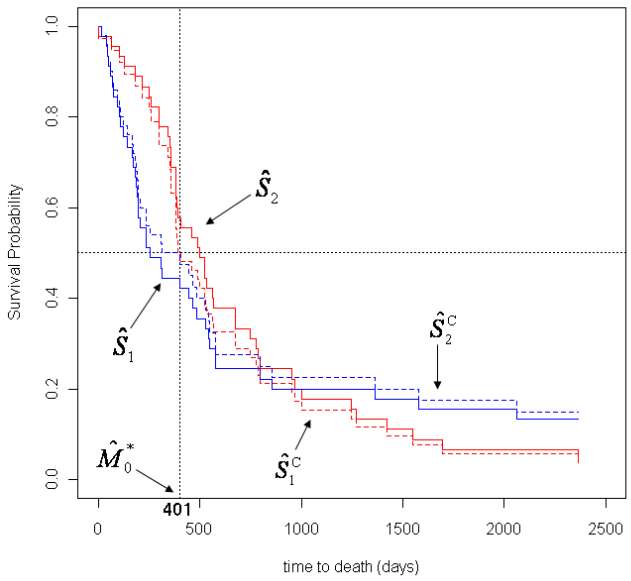

Figure 30: Kaplan-Meier estimates $\widehat{S}_j(t)$ and the constraint Kaplan-Meier estimates $\widehat{S}_j^c(t)$, $j = 1, 2$ (dotted line) for gastric cancer data

5. Conclusion and Further Works

In this thesis, we discussed a testing procedure in order to detect the difference of medians in the framework of two-sample problem including right censoring. To refer existing median tests for right censored data, we introduced three existing median tests. First, the generalized sign test has been developed by Brookmeyer and Crowley (1982b) which is based on the extension of sign test to censored data version. Since their testing procedure is based on the asymptotic normality of Kaplan-Meier estimate, it can be considered to have small power in particular situation, namely, in small samples. Also, this testing procedure does not provide the same test statistic value for the replacement of information between two samples. In order to avoid the confusion caused by different two p-values, we defined the unified test statistic T'_{BC}. Second, Naik-Nimbalkar and Rajarshi (1997) proposed the empirical likelihood ratio test for the equality of k medians in right censored data. One of the key characteristics of their testing procedure is that it is based on $\mathrm{H}_0^{\mathrm{m}}$ while other median tests are based on H_0. Although their idea is considered a natural way for constructing median test based on the empirical profile likelihood, they did not show any numerical investigation in their paper. So, we introduced a practical way until the derivation of p-value by using the Newton-Raphson method where the Lagrangian parameters are estimated in order to find the estimate of the common median ($\widehat{M}_0^*$) under $\mathrm{H}_0^{\mathrm{m}}$. Using the estimated Lagrangian parameters and common median with maximum profile log likelihood, the constrained maximum log likelihood value can be calculated. Thus, one can obtain the empirical likelihood ratio test statistic based on the unconstrained maximum log likelihood value and the constrained maximum log likelihood value. Last, Park and Na (2000) proposed a bootstrap median test for right-censored data. Although their bootstrap median test is rather simple to estimate the bootstrap p-value, it requires a heavy workload on the computation of p-value. Based on several unfavorable char-

acteristics of existing median tests for right censored data, we proposed an alternative median test based on the two-sample difference between the mid order statistics, and the significance probability function. Furthermore, we provided the rationale to estimate the significance probability, a manner to cope with censored observations and some contrivances to overcome computational problem in that estimation. Although this proposed median test was developed for testing H_0 mainly, we also introduced a modified version of proposed median test for testing $\mathrm{H}_0^{\mathrm{m}}$ by making use of the constrained Kaplan-Meier estimates. A proposal for an alternative median test for right censored data was one of the main objectives of this thesis. To investigate the performances of the proposed median test, simulation and case studies have been carried out.

We investigated the null distribution, the empirical type I error and power for each of median tests under several hypotheses based on simulation studies. From these simulation studies, we found that the null distributions of $\widehat{pv}(x)$ and $\widehat{p}_{boot}$ are almost uniform except both tails. On the other hand, we found that the null distributions of T'_{BC} and T_{ELR} are not uniform. In particular, it was suggested that null distribution of T_{ELR} is not continuous for small samples. We have no clear reasons why it has such distribution, so it is valuable to pursue the exact cause of such trend for T_{ELR} in future. With respect to the empirical type I error, only T'_{BC} provided the error rate of greater than 0.05 regardless of sample size and percentage of censoring. From these investigations, it was suggested that the null distributions of $\widehat{pv}(x)$ and $\widehat{p}_{boot}$ are asymptotically valid, and they are appropriate with a conservative tendency in the finite sample. T_{ELR} may be appropriate in such sense that it had conservative tendency of the empirical type I error. However, as described above, we will need to find the reason for the unique shape of null distribution. For T'_{BC} based median test, it was suggested that its null distribution may not be valid, and their empirical type I error was extremely higher for small samples.

With respect to the power, it was suggested that $\widehat{pv}(x)$ has higher power than that of T'_{BC} under the simulation model in which the discrepancy of two survival distributions is increased over time (Simulation study 1). We think that $\widehat{pv}(x)$ had higher power because of it property to consider survival information after medians. In contrast, it can be considered that T'_{BC} had lower power because it must lose the survival information after medians, that is, it does not consider the discrepancy between two survival distributions

post medians. Second simulation study provided powers for each of median tests under shift model as well. Unlike the first simulation study, the discrepancy between two survival distributions is fixed. Because of such characteristic of alternative hypothesis, $\widehat{pv}(x)$ had smaller power than that of T'_{BC} for small samples. However, $\widehat{pv}(x)$ had the same or higher power than that of T'_{BC} for each of underlying survival distributions on average in large samples. From this, it was suggested that $\widehat{pv}(x)$ tends to be powerful as increasing n compared to T'_{BC} under shift model. T_{ELR} had lower power than $\widehat{pv}(x)$ for small samples. As the empirical type I error of T_{ELR} was extremely small, so it is expected to lead such results. Bootstrap median test revealed that it has the same power as others in larger sample while it had the smallest power for small samples (under shift model).

We discussed the nonparametric Behrens-Fisher problem by distinguishing between H_0 and $\mathrm{H}_0^{\mathrm{m}}$ in the beginning of section 2. As described in background of our thesis, our interest is to detect the difference of survival distributions based on the median survival time. Therefore, our main interest is to test H_0. However, one may be interested in $\mathrm{H}_0^{\mathrm{m}}$, namely, $\{M_1^* = M_2^* \text{ and } S_1(t) \neq S_2(t)\}$, so we modified a proposed median test to test $\mathrm{H}_0^{\mathrm{m}}$ where $\widehat{pv}(x)$ is modified by replacing $\widehat{S}_0$ by the constrained Kaplan-Meier estimates $\widehat{S}_j^{\mathrm{c}}$ $(j = 1, 2)$ which were discussed in the empirical likelihood ratio test. As one of the further works, it is important to investigate the difference of performance between H_0 and $\mathrm{H}_0^{\mathrm{m}}$ in detail and the correlation structure between them by focusing on $\{M_1^* = M_2^*$ and $S_1(t) \neq S_2(t)\}$. Also, one of the important perspectives to be noted in the survival study is to shorten the duration of clinical trial. Considering that main goal, we need to compare median tests with existing rank tests such as log-rank test and generalized Wilcoxon test in detail from a view point of early detection on the clinical trial result.

In this thesis, we used Kaplan-Meier estimate to estimate the significance probability function. However, we are also interested in parametric inference to compare medians based on the significance probability function. Especially, we are interested in the inference of the significance probability function applied to the normal distribution family so that we could compare median test with t-test and/or Wilcoxon-test.

Appendix

A.1 Density for the difference in median survival times: Asymptotic result by Laplace approximation

Here, we discuss a point based on the Laplace approximation to derive the limiting form for the density of the difference in median survival times. However, since the same limiting distribution can be derived by another theory, we provide no detailed discussions considering the rest term and order for the convergence. Asymptotic theory based on the Laplace approximation provides several useful suggestions in discussing the property of the proposed median test.

By Stirling's approximation for Gamma function, Beta density $-\phi_j(x)dS_j(x)/\mathrm{B}(m_j, m_j)$ is asymptotically equivalent to the normal density

$$-\left(1/\sqrt{2\pi}\sigma(n_j)\right)\exp\left\{-(S_j(x)-0.5)^2/2\sigma(n_j)^2\right\}dS_j(x) \tag{A.1}$$

where, $\sigma(n_j)=\sqrt{1/4n_j}$ $\left(\sigma(n_j)=\sqrt{1/4n'_j}\text{ for censored data}\right)$. And, the density function $g_1(y)$ in (2.12) is asymptotically equivalent to

$$\bar{g}_1(y)=\frac{1}{2\pi\sigma(n_1)\sigma(n_2)}\int_y^\infty \exp\left\{-\frac{(S_1(v)-0.5)^2}{2\sigma(n_1)^2}-\frac{(S_2(v-y)-0.5)^2}{2\sigma(n_2)^2}\right\}f_2(v-y)f_1(v)dv.$$

Hence, we assume H_0 where $S_0(t)=S_1(t)=S_2(t)$ and $M_0^*=S_0^{-1}(0.5)$ as well as used in the context. Furthermore, to simplify the remaining discussion, we define $v'=v-\kappa y$ for a constant $\kappa(>0)$, and then we perform variable transformation with respect to v of $\bar{g}_1(y)$. We discuss how to deal with κ later. Through such variable transformation (and we bring v' back to v), $\bar{g}_1(y)$ is

$$\bar{g}_1(y)=\frac{1}{2\pi\sigma(n_1)\sigma(n_2)}\int_{y\kappa'}^\infty \exp\left[h(v;y,\kappa)\right]f_0(v-\kappa' y)f_0(v+\kappa y)dv$$

where $\kappa'=1-\kappa$, and

$$h(v;y,\kappa)=-\{S_0(v+\kappa y)-0.5\}^2/2\sigma(n_1)^2-\{S_0(v-\kappa' y)-0.5\}^2/2\sigma(n_2)^2.$$

By differentiating $h(v; y, \kappa)$ with respect to v twice, then the first-order differential $h_v(v; y, \kappa)$ and the second-order differential $h_{vv}(v; y, \kappa)$ are

$$
\begin{aligned}
h_v(v; y, \kappa) &= (S_0(v + \kappa y) - 0.5) f_0(v + \kappa y)/\sigma(n_1)^2 + (S_0(v - \kappa' y) - 0.5) f_0(v - \kappa' y)/\sigma(n_2)^2 \\
h_{vv}(v; y, \kappa) &= \{-f_0(v + \kappa y)^2 + (S_0(v + \kappa y) - 0.5) f_0'(v + \kappa y)\}/\sigma(n_1)^2 \\
&\quad + \{-f_0(v - \kappa' y)^2 + (S_0(v - \kappa' y) - 0.5) f_0'(v - \kappa' y)\}/\sigma(n_2)^2
\end{aligned}
$$

respectively, where $f_0'(x) = \partial f_0(x)/\partial x$. Function $h(v; y, \kappa)$ has a maximum point at $v = \widehat{v}$ which satisfy $h_v(\widehat{v}; y, \kappa) = 0$. Such $\widehat{v}$ arises in the neighborhood of M_0^* (when $S_0(\widehat{v} + \kappa y)$ and $S_0(\widehat{v} - \kappa' y)$ are near 0.5 together) from the shape of $h_v(v; y, \kappa)$, and M_0^* is an inner point between $\widehat{v} - \kappa' y$ and $\widehat{v} + \kappa y$, that is, $\widehat{v} - \kappa' y \leq M_0^* \leq \widehat{v} + \kappa y$. So, one can set $\widehat{v} + \kappa y - M_0^* = \kappa' y$ for any κ, then the followings can be obtained by the Taylor expansion,

$$
\begin{aligned}
S_0(\widehat{v} + \kappa y) &= S_0(M_0^*) - f_0(M_0^*)\kappa' y - f_0'(\tilde{M}_{10})\kappa'^2 y^2/2, \quad (\exists \tilde{M}_{10} \in (M_0^*, \widehat{v} + \kappa y)), \\
S_0(\widehat{v} - \kappa' y) &= S_0(M_0^*) + f_0(M_0^*)\kappa y - f_0'(\tilde{M}_{20})\kappa^2 y^2/2, \quad (\exists \tilde{M}_{20} \in (\widehat{v} - \kappa' y, M_0^*)),
\end{aligned}
\tag{A.2}
$$

where $S_0(M_0^*) = 0.5$. In addition, we obtain the following relationship to determine κ by substituting (A.2) into $h_v(\widehat{v}; y, \kappa) = 0$:

$$
\frac{f_0(M_0^*)\kappa - f_0'(\tilde{M}_{20})\kappa^2 y/2}{f_0(M_0^*)\kappa' + f_0'(\tilde{M}_{10})\kappa'^2 y/2} = \frac{\sigma(n_2)^2}{\sigma(n_1)^2} \frac{f_0(\widehat{v} + \kappa y)}{f_0(\widehat{v} - \kappa' y)}. \tag{A.3}
$$

Thus, three quantities κ, $S_0(\widehat{v} + \kappa y) - 0.5$ and $S_0(\widehat{v} - \kappa' y) - 0.5$ can be determined by (A.2) and (A.3). Since the integral in $\bar{g}_1(y)$ is dominant at a point $v = \bar{v}$ (around the neighborhood of M_0^*) which provides a maximum point of $h(v; y, \kappa)$, the results of Laplace approximation provides that $\bar{g}_1(y)$ is asymptotically equivalent to

$$
\bar{\bar{g}}_1(y) = \frac{f_0(\widehat{v} - \kappa' y) f_0(\widehat{v} + \kappa y)}{\sqrt{2\pi}\sigma(n_1)\sigma(n_2)\sqrt{-h_{vv}(\widehat{v}; y, \kappa)}} \exp\{h(\widehat{v}; y, \kappa)\}
$$

From $y \approx O_p(1/\sqrt{n})$ by (A.1), (A.2) and (A.2), we finally conclude that $\bar{\bar{g}}_1(y)$ is asymptotically equivalent to (2.14) (asymptotically equivalent to (2.17) for censored data).

A.2 Evaluation of $\mathrm{E}[\widehat{pv}_{13}]$

$\widehat{pv}_{13}(x)$ is written by $\widehat{pv}_{13}(x) = B_{m_1 m_2} \int_0^{\tau_n - x} \{\int_{t+x}^{\tau_n} d\widetilde{\mathcal{A}}_1(s)\} d\widetilde{\mathcal{M}}_2(t)$ based on the exchange order of integration. In the finite sample, integrand $\int_{t+x}^{\tau_n} d\widetilde{\mathcal{A}}_1(s)$ is not $\mathscr{F}_t$-predictable. However, we have $\mathrm{E}[\widehat{pv}_{13}(x)|_{\widetilde{\mathcal{A}}_1 = \mathcal{A}_1^*}] = 0$ by the martingale property when $\widetilde{\mathcal{A}}_1$ is replaced by definitive $\mathcal{A}_1^*$, so

$$\mathrm{E}[\widehat{pv}_{13}(x)] = B_{m_1 m_2} \mathrm{E}[\textstyle\int_0^{\tau_n - x} \int_{t+x}^{\tau_n} (d\widetilde{\mathcal{A}}_1(s) - d\mathcal{A}_1^*(s)), d\widetilde{\mathcal{M}}_2(t)],$$

where $d\mathcal{A}_j^*(s) = \phi_j(s_-) S_0(s_-) d\Lambda_0(s)$. First, for simplicity, we consider a case where $B_{m_1 m_2}$ and the power m_j in $\widehat{\phi}_j(\cdot)$ are fixed. Here, we have $d\widetilde{\mathcal{A}}_1(s) - d\mathcal{A}_1^*(s) = O_p(n^{-1/2}) d\Lambda_0(s)$ by the asymptotic normality of $\sqrt{n}(\widehat{S}_0(s) - S_0(s))$ and delta method. In addition, $n\mathrm{E}[|\widehat{pv}_{13}(x)|] < \infty$ can be shown by a result $n\langle \widetilde{\mathcal{M}}_j \rangle(\tau) = O_p(1)$ given in the proof of Theorem 1. So, we obtain $n\mathrm{E}[\widehat{pv}_{13}(x)] \to_p 0$ by dominated convergence theorem and asymptotic martingale properties of $\widehat{pv}_{13}(x)$, that is, $\mathrm{E}[\widehat{pv}_{13}(x)] = o(n^{-1})$. We show how such results can be extended in usual case where $m_1, m_2 \to \infty$ as well as n. As $\bar{g}_1(y)$ in Appendix A.1, $\widehat{pv}_{13}(x)$ is equivalent to the following expression

$$-\frac{\sqrt{\eta_1 \eta_2}}{\pi} \int_x^{\tau_n} \int_y^{\tau_n} \exp\left\{-\eta_2 \widehat{Z}_{\mathrm{m}}(v - y_-)^2\right\} \widehat{C}(v_-) \frac{\widehat{S}_0(v - y_-)}{S_0(v)} \frac{d\overline{\mathcal{M}}_{\mathrm{m}}(v - y)}{\overline{\mathcal{Y}}(v - y)} dZ_{\mathrm{m}}(v), \quad \text{(A.4)}$$

by Stirling's approximation, where $\eta_j = 2n_j/n$, $\widehat{Z}_{\mathrm{m}}(v) = \sqrt{n}(\widehat{S}_0(v) - 0.5)$, $Z_{\mathrm{m}}(v) = \sqrt{n}(S_0(v) - 0.5)$, $\overline{\mathcal{M}}_{\mathrm{m}}(v - y) = \sqrt{n}(\overline{\mathcal{M}}(v - y) - \overline{\mathcal{M}}(M_0^*))$, and

$$\widehat{C}(v_-) = \exp\{-\eta_1 \widehat{Z}_{\mathrm{m}}(v_-)^2\} \widehat{S}_0(v_-) - \exp\{-\eta_1 Z_{\mathrm{m}}(v_-)^2\} S_0(v_-).$$

Under situation where (A.4) is obtained, $(\widehat{S}_0(v) - 0.5)/(v - M_0^*) \to_p (S_0(v) - 0.5)/(v - M_0^*)$ since v is concentrated in the neighborhood of M_0^* with order $(v - M_0^*) \cong O_p(1/\sqrt{n})$ stochastically, so $\widehat{Z}_{\mathrm{m}}(v) \to_p Z_{\mathrm{m}}(v)$. As a result, we obtain $n^{-1/2} \widehat{C}(v_-) = O_p(1)$. By similar discussions, $d\langle \overline{\mathcal{M}}_{\mathrm{m}} \rangle(v - y) = O_p(n^{1/2}) \overline{\mathcal{Y}}(v - y)\, d\Lambda_0(v - y)$, and $d\langle \overline{\mathcal{M}}_{\mathrm{m}} \rangle(v - y)/\overline{\mathcal{Y}}(v - y)^2 = O_p(n^{-1}) dZ_{\mathrm{m}}(v - y)$. Once we apply these results to (A.4), we have $n\mathrm{E}[|\widehat{pv}_{13}(x)|] < \infty$, and obtain $\mathrm{E}[\widehat{pv}_{13}(x)] = o(n^{-1})$ definitely.

References

1. Anderson, J. R., Bernstein, L. and Pike, M. C. (1982). Approximate confidence intervals for probabilities of survival and quantiles in life-table analysis. *Biometrics*, **38**, 407-416.

2. Andersen, P. K., Borgan, Ø., Gill, R. D. Keiding, N. (1993). *Statistical models based on counting processes.* Springer-Verlag.

3. Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course in Order Statistics.* John Wiley and Sons.

4. Bartholomew, D. J. (1957). A problem in life testing. *Journal of the American Statistical Association*, **52**, 350-355

5. Bie, O., Borgan, O., and Liestol, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small properties. *Scandinavian Journal of Statistics*, **14**, 221-233.

6. Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics*, **2**, 437-453.

7. Brookmeyer, R. and Crowley, J. J. (1982a). A confidence interval for the median survival time. *Biometrics*, **38**, 29-41.

8. Brookmeyer, R. and Crowley, J. J. (1982b). A k-sample median test for censored data. *Journal of the American Statistical Association*, **77**, 433-440.

9. Chap T. LE. (1997). *Applied Survival Analysis.* John Wiley and Sons.

10. Collett, D. (1994). *Modeling Survival Data in Medical Research.* New York:Champman and Hall.

11. Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall.

12. Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall.

13. Desu, M. M. and Raghavarao, D. (2004). *Nonparametric statistical methods for complete and censored data.* Chapman and Hall.

14. Efron, B. (1967). The two-sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol.* 4, L. LeCam and J. Neyman(eds), 831-854. Berkeley ; University of California Press.

15. Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, **76**, 312-319.

16. Emerson, J. (1982). Nonparametric confidence intervals for the median in the presence of right censoring. *Biometrics*, **38**, 17-27.

17. Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* John Wiley and Sons.

18. Gastwirth, J. and Wang, J. L. (1988). Control percentile test procedures for censored data. *Journal of statistical planning and inference*, **18**, 267-276.

19. Gehan, E. A. (1965). A generalized wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*, **52**, 203-223.

20. Hájek, J. and Sidák, Z. (1967). *Theory of rank tests.* Academia, Prague and Academic Press.

21. Jennison, C. and Turnbull, B. W. (1985). Repeated confidence intervals for the median survival time. *Biometrika*, **72**, 619-625.

22. Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* John Willey and Sons.

23. Kaplan, E. L. and Meier. P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

24. Karrison, T. (1996). Confidence intervals for median survival times under a piecewise exponential model with proportional hazards covariate effects. *Statistics in medicine*, **15**, 171-182.

25. Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis*: *Techniques for Censored and Truncated Data*, 2nd edition. Springer-Verlag.

26. Lagakos, S. W. (1979). General right censoring and its impact on the analysis of survival data. *Biometrics*, **35**, 139-156.

27. Latta, R. B. (1981). A monte carlo study of some two-sample rank tests with censored data. *Journal of the American Statistical Association*, **76**, 713-719.

28. Naik-Nimbalkar, U. V. and Rajarshi, M. B. (1997). Empirical likelihood ratio test for equality of k medians in censored data. *Statistics & Probability Letters*, **34**, 267-273.

29. Park, H. I. and Na, J. H. (2000). Bootstrap Median Tests for Right Censored Data. *Journal of the Korean Statistical Society*, **29**, 423-433.

30. Reid, N. (1981). Estimating the median survival time. *Biometrika*, **68**, 601-608.

31. Sickle-Santanello, B. J., Farrar, W. B., Keyhani-Rofagha, S., Klein, J. P., Pearl, D., Laufman, H., Dobson, J., and O'Toole, R. V. (1988). A Reproducible System of Flow Cyto metric DNA Analysis of Paraffin Embedded Solid Tumors: Technical Improvements and Statistical Analysis. *Cytometry*, **9**, 594-599.

32. Simon, R. and Lee, Y. J. (1982). Nonparametric confidence limits for survival probabilities and median survival time. *Cancer Treatment Reports*, **66**, 37-42.

33. Slud, E. V., Byar, D. P. and Green, S. B. (1984). A comparison of reflected versus test-based confidence intervals for the median survival time, based on censored data. *Biometrics*, **40**, 587-600.

34. Stablein, D. M. and Koutrouvelis I. A. (1985). A two-sample test sensitive to crossing hazards in uncensored and singly censored data. *Biometrics*, **41**, 643-652.

35. Su, J. Q. and Wei, L. J. (1993). Nonparametric estimation for the difference or ratio of median failure times. *Biometrics*, **49**, 603-607.

36. Wang, J. L. and Hettmansperger, T. P. (1990). Two-sample inference for median survival times based on one-sample procedures for censored survival data. *Journal of the American Statistical Association*, **85**, 529-536.

List of Publications

Amagasaki, T., Sugimoto, T., Matsubara, Y. and Goto, M. (2000). Comparative inference of two samples based on median survival (in Japanese). *Journal of the Society for Clinical and Biostatistical Research on Cancer*, **21**, 19–24.

Amagasaki, T. (2001). Parametric inference for median survival time in two samples. *Master thesis*, Osaka University.

Amagasaki, T., Sugimoto, T., Matsubara, Y. and Goto, M. (2001). Parametric inference for median survival time in two samples. *Proceedings of the 15th Conference of the Japanese Society of Computational Statistics*, Okayama, Japan.

Amagasaki, T., Sugimoto, T. and Goto, M. (2006). Nonparametric inference for median survival time in two samples. *Proceedings of the 20th Symposium of the Japanese Society of Computational Statistics*, Tokyo, Japan.

Amagasaki, T., Sugimoto, T. and Goto, M. (2009). An comparative inference in two samples based on the median survival time (in Japanese). *Journal of the Japan Statistical Society*, **39**, 95–119.