指数分布に従うデータへの層別ログランク検定の検出力の検討

○小松 邦岳（株式会社アスクレップ）

（臨床評価研究会 基礎解析分科会）

Impact of power of the stratified Log-Rank test for survival data based on exponential distribution.

Kunitake Komatsu（ASKLEP Inc.）
The association of clinical evaluation

要旨

層別生存時間解析とは、生存時間解析で比較する群の中に生存関数に違いのある因子が認められた場合、その因子で層別し個別に生存時間解析を実施、それぞれの結果を併合する手法である。層別を行わなかった場合に起こる検出力の低下を、層別を行うことにより防ぐことができるが、状況によっては層別を行うことにより逆に検出力が低下してしまう。

この演題は、さまざまな指数分布に従った乱数を用いたシミュレーションによって、層別を行った場合と行わなかった場合の検出力の違いを示すことを目的とする。

キーワード：層別生存時間解析、層別ログランク検定、検出力

1 はじめに

ログランク検定は、症例が不均一ではないことを仮定しており、症例が不均一な場合には検出力が低下する[1]。症例が不均一とは予後に影響のある因子、すなわち生存関数に影響を与える因子が症例ごとに異なって存在していることである。ある因子が生存関数に影響を与えているとき、その因子で層別を行ったログランク検定により、層間の生存関数の違いを認めたうえでの解析結果を得ることができ、層別を行わなかった場合と比べ、検出力の低下を抑止できる[1]。ただし、層間には治療効果が一様であること、つまり治療効果と層別因子間に交互作用がないことが前提となる[2]。ではどのような場合にどの程度検出力は変化するのだろうか。

指数分布に従った乱数を用いて作成した生存時間データのパラメータをさまざまに変化させ、層別を行わないログランク検定、層別を行ったログランク検定のシミュレーションを実施し、検出力の違いを検討した。

なお、以降本稿中では層別を行わないログランク検定を「単純ログランク検定」、層別を行うログランク検定を「層別ログランク検定」と呼称する。また、シミュレーションデータは全て 2 群、2 層とした。

2 層別ログランク検定について

層別ログランク検定の実施方法を以下に紹介する。なお、ログランク検定そのものについてはここでは触れない。

2.1 層別ログランク検定の計算方法

手順としては、それぞれの層ごとにスコア統計量 u およびその分散成分 V を求め、それを元にして併合した統計量（χ^2値）を算出する。併合した統計量は自由度［層の数-1］のχ^2分布に従うことを利用し統計的仮説検定を行う。

層ごとのスコア統計量 u およびその分散成分 V は、層で分割したデータからそれぞれに算出すればよい。2 群の場合の各層における算出方法は以下のとおりである[3]。

スコア統計量：

$$u_j = \sum_i (d_{ij} - e_{ij})$$

スコア統計量の分散成分：

$$V_j = \sum_i \frac{(n_i - n_{ij}) n_{ij} d_i (n_i - d_i)}{{n_i}^2 (n_i - 1)}$$

- e_{ij} ：イベント数の期待値
- d_{ij} ：時点 i におけるイベント数
- n_{ij} ：時点 i における j 群のリスク集合の大きさ
- n_i ：時点 i における群を併合したリスク集合の大きさ

S 個の層ごとに算出したスコア統計量 u および V から、併合した統計量（${\chi^2}_t$）は以下のように求める[2]。

$${\chi^2}_t = \left(\sum_{k=1}^{S} u_k\right)^2 \Big/ \left(\sum_{k=1}^{S} V_k\right)$$

- ${\chi^2}_t$：併合した統計量
- u_k ：層ごとに算出したスコア統計量 u
- V_k ：層ごとに算出したスコア統計量の分散成分 V

この統計量（${\chi^2}_t$）を自由度［層の数-1］の χ^2 乗分布の任意の%点と比較することにより、有意か否かを得ることができる。

2.2 SAS program

lifetest プロシジャによる層別ログランク検定実施のプログラム例を以下の通り提示する[4][5]。

SAS program2.2.1：lifetest プロシジャによる層別ログランク検定例

```
proc lifetest data=入力データセット名 ;
   time 時間変数 * 打ち切り変数(打ち切り値) ;
   strata 層別変数 / group=群別変数 ;   run;
```

3 シミュレーション方法

3.1 概要・手順

指数分布に従った生存時間データの乱数データセットを 4 つ生成し、縦結合をした 1 つのデータセットについて単純ログランク検定および層別ログランク検定を実施する。これを繰り返し、検出力の算出および検討を行う。

4 つの乱数データセットは群間の差、層別の差を考慮しハザードなどのパラメータを変化させる。結合した乱数データに対して単純ログランク検定と層別ログランク検定を実施し、それぞれの検定結果で有意差の有

無を確認する。これを、乱数データセットを変えながら複数回繰り返したときに有意となった割合が検出力となる。検討のため、検出力の比（単純ログランク検定の検出力÷層別ログランク検定の検出力）も合わせて算出する。特に断りが無い限りシミュレーションの回数は 1 万回とした。

なお、データおよびシミュレーションは特定の疾患や状況を仮定していないことを断っておく。

3.2 指数分布に従う乱数列の生成方法

指数分布に従う生存時間データを、以下の方法で生成した[1]。

指数分布に従う生存関数は$S(t) = exp[-\lambda t]$と表され、これをイベント発現までの期間 t について解くと以下の数式が得られる。

$$t = -\frac{\log(S(t))}{\lambda} \qquad \text{（式 3.2.1）}$$

t　　：生存時間
$S(t)$：生存関数
λ　：ハザード

式 3.2.1 に対して、λに任意の値、$S(t)$に 0～1 の値をとる一様乱数を与えることにより、指数分布モデルに従ったシミュレーションデータとなるイベント発現時期tを得た。

3.3 指数分布に従う乱数列の生成方法（SAS プログラム）

乱数列発生の SAS プログラム例を以下に示す。このプログラムでは、任意の時点（マクロパラメータ：CNS）以降は全てその時点で打ち切りとしている。

SAS program3.3 ： 指数分布に従う生存時間シミュレーションデータ作成 SAS プログラム例

```
%macro DATAGEN(LAMBDA=,N=,CNS=);
  do I=1 to &N ;
    ST = rand("uniform");
    TIME = ((-1)*(log(ST)))/(&LAMBDA) ;
    CENSOR=0 ;
    if(TIME>&CNS)then do;
      CENSOR=1 ;
      TIME = &CNS ;
    end;  output;  end;
%mend DATAGEN ;
```

・マクロパラメータ
　LAMBDA ＝ハザード（λ）を設定。
　N＝生成するレコード数を設定。
　CNS＝打ち切り時点を設定。

3.4 層別生存時間解析のシミュレーションデータセットの生成方法

指数分布に従う乱数を 4 種類生成し、それぞれを群・層とした 1 つのデータセットとして縦結合する。基本となるデータセットを 3.5 の通り定め、そこから様々に乱数発生のパラメータを変化させ、それぞれの場合での検出力をシミュレーションにより導出、検討する。

3.5 基本となる乱数データセット

基本となる乱数データセットは以下のように定めた。

	層 A	層 B
群 1	例数：N_{1A}＝75 ハザード：λ_{1A}=0.6	例数：N_{1B} ＝75 ハザード：λ_{1B}=0.9
群 2	例数：N_{2A} ＝75 ハザード：λ_{2A}=0.84	例数：N_{2B} ＝75 ハザード：λ_{2B}=1.26
観察期間：t=4 で打ち切り。		

このデータセットは以下のような特徴を持たせている。

- 層 A、B による層別ログランク検定の検出力が約 0.8 となる。（10 万回のシミュレーションによる検出力計算を 10 回繰り返した結果、検出力の平均±標準偏差は 0.8092±0.0015。）
- 群、層において症例数が同数である。
- 各群に共通で層 A、層 B 間のハザードの比が 1.5 倍。
- 各層に共通で群 1、群 2 間のハザードの比が 1.4 倍。
- このデータセットと同じ分布に従う生存曲線について、1 層あたり 10 万症例としてシミュレーションした結果のグラフを、理論曲線に換えて示す（打ち切りはしない）。群ごと、層ごとに示した。また、二重対数変換したプロットも合わせて示した。

図4. 1. 1 基本となるデータの生存曲線（群別）

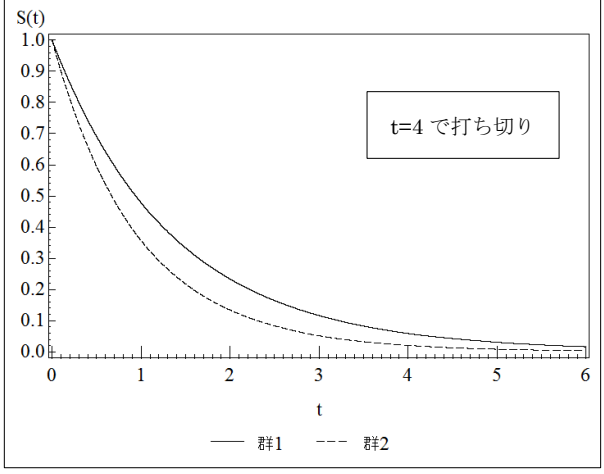


図4. 1. 2 基本となるデータの生存曲線：二重対数変換（群別）

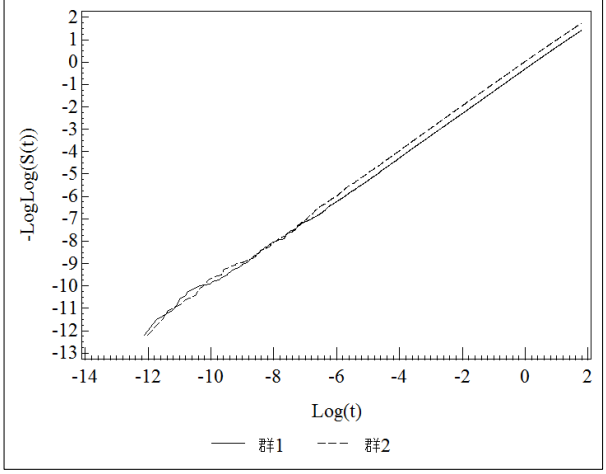


図4. 1. 3 基本となるデータの生存曲線（層別）

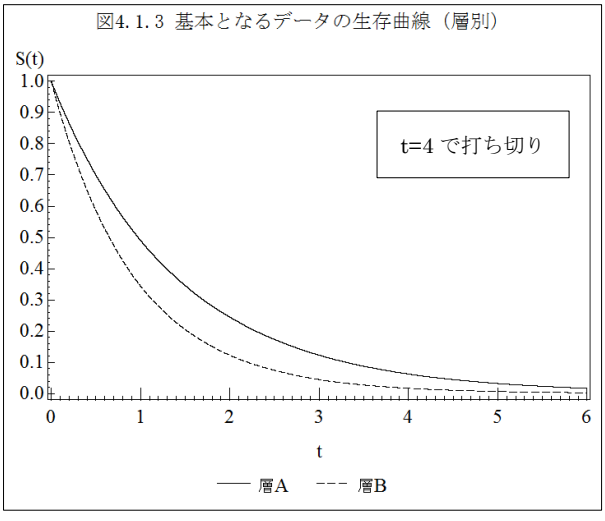
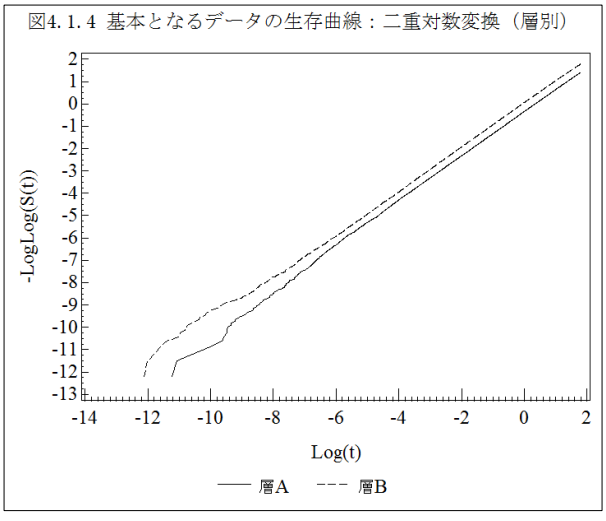


図4. 1. 4 基本となるデータの生存曲線：二重対数変換（層別）



4 シミュレーション結果

4.0 シミュレーションを行った検討項目

シミュレーションによって、以下の内容を検討した。

4.1.症例数の変化に伴う検出力の検討

4.2.観察期間の変化に伴う検出力の検討

4.3.層間のハザードの変化に伴う検出力の検討

4.4.1 層間の例数がアンバランスな時の検出力の検討(層内ではバランス)

4.4.2 層間の例数がアンバランスな時の検出力の検討(層内でアンバランス)

4.5.層間の一様性が仮定できないときの検出力の検討

4.1 症例数の変化に伴う検出力の検討

1 層あたりの症例数を変化させ、検出力の推移を確認した。その際、各群・層の症例数は同数とした。

【データの説明】

	層 A	層 B
群 1	例数：N_{1A} ハザード：λ_{1A}=0.6	例数：N_{1B} ハザード：λ_{1B}=0.9
群 2	例数：N_{2A} ハザード：λ_{2A}=0.84	例数：N_{2B} ハザード：λ_{2B}=1.26
t=4 で打ち切り。 N_{1A} =N_{1B} =N_{2A} =N_{2B}=5, 10, 25, 50, 75, 100, 125, 150, 200		

【シミュレーション結果と検討】

結果を図 4.1.1 に示した。全体の傾向として、例数が増えるほど検出力が向上した。

症例数が 25～125 の間で層別ログランク検定よりも単純ログランク検定の検出力が低かった。それ以外の症例数では、検出力が過大（1 に近い）または検出力が過小（0 に近い）の場合となり、検出力の違いは確認できなかった。

このデータでは双方の検出力が拮抗していたため、さらに λ_{1B}= 1.2、λ_{2B}= 1.68 とした場合の検出力についても示す（図 4.1.2）。

こちらについても、同様の傾向が見られたが、層別を行った場合と行わない場合の検出力の違いについてはより顕著となった。1 層あたりの症例数が 25 の場合には層別を行った場合の検出力が 0.35 なのに対し、層別を行わない検出力は 0.31 であり、検出力の比は 0.87 倍と低下している。層別を行った場合の検出力が約 0.8 となる 1 層当たりの症例数が 75 の場合でも、層別を行わない検出力は 0.74 となり、0.92 倍となった。

図4.1.1 症例数の変化による検出力の検討

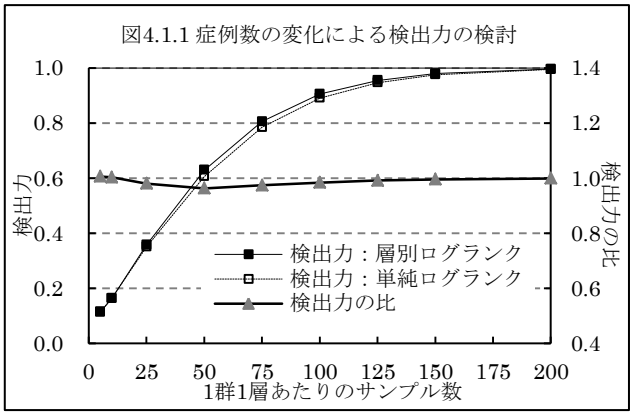
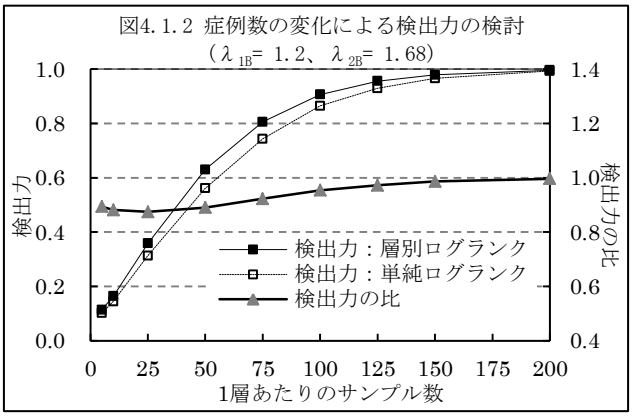


図4.1.2 症例数の変化による検出力の検討
（λ_{1B}= 1.2、λ_{2B}= 1.68）



4.2 観察期間の変化に伴う検出力の検討

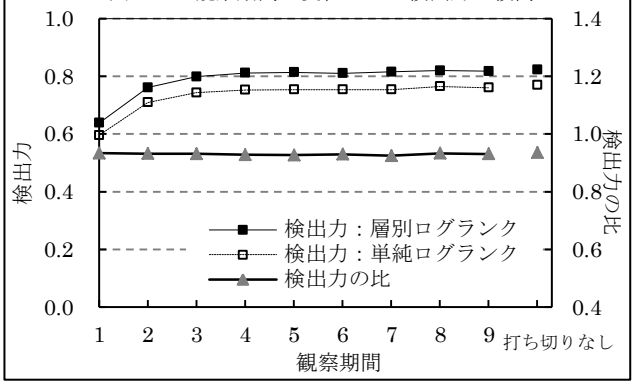
観察期間を変化させ、検出力の推移を確認した。

【データの説明】

	層 A	層 B
群 1	例数：N_{1A}＝75 ハザード：λ_{1A}=0.6	例数：N_{1B} ＝75 ハザード：λ_{1B}=0.9
群 2	例数：N_{2A} ＝75 ハザード：λ_{2A}=0.84	例数：N_{2B} ＝75 ハザード：λ_{2B}=1.26
観察期間 : t=1,2,3,4,5,6,7,8,9,10 で打ち切り。および打ち切りなし。		

【シミュレーション結果と検討】

図4. 2. 1 観察期間の変化による検出力の検討



結果を図 4.2.1 に示した。全体の傾向として、観察期間が非常に短い場合において検出力が低下する様子が見られる。観察期間=4 以降において、検出力は一定した。

層別を行った場合と行わない場合の検出力の比較については、層別を行わないことによる検出力の低下が見られたが、検出力の比は観察期間の設定を問わず約 0.93 倍で一定であった。

4.3 層間のハザードの変化に伴う検出力の検討

層 B のハザードを変化させ、検出力の推移を確認した。

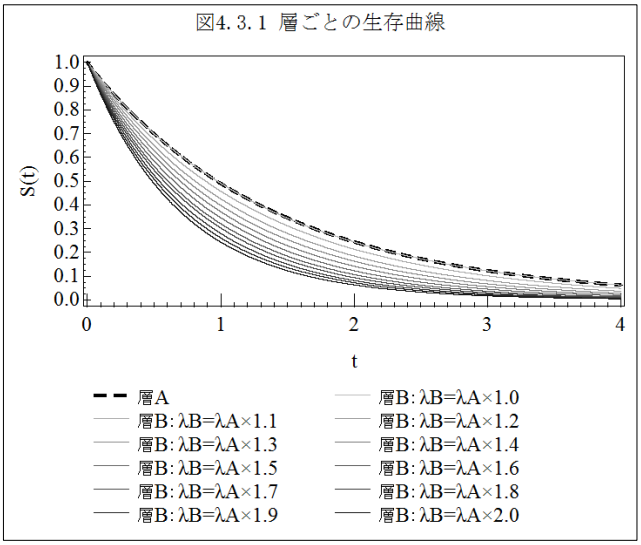
【データの説明】

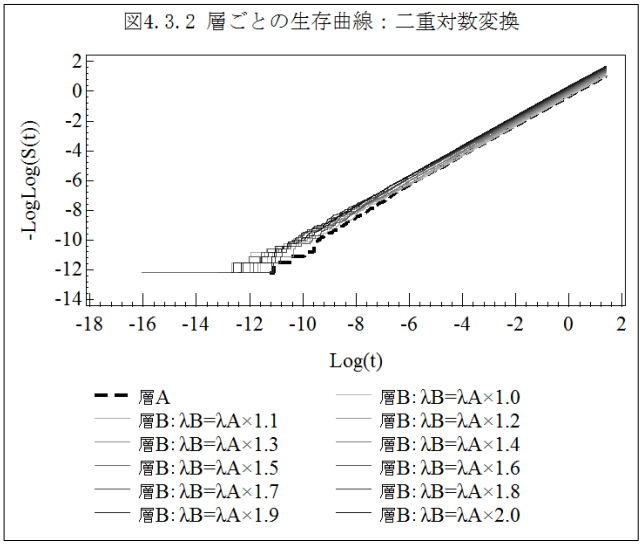
	層 A	層 B
群 1	例数：N_{1A}＝75 ハザード：λ_{1A}=0.6	例数：N_{1B} ＝75 ハザード：λ_{1B} 下記
群 2	例数：N_{2A} ＝75 ハザード：λ_{2A}=0.84	例数：N_{2B} ＝75 ハザード：λ_{2B} 下記
観察期間：t=5 で打ち切り。 λ_{1B}, λ_{2B}＝λ_{1A}、λ_{2A} の 1 倍,1.1 倍,1.2 倍,…,1.9 倍,2.0 倍		

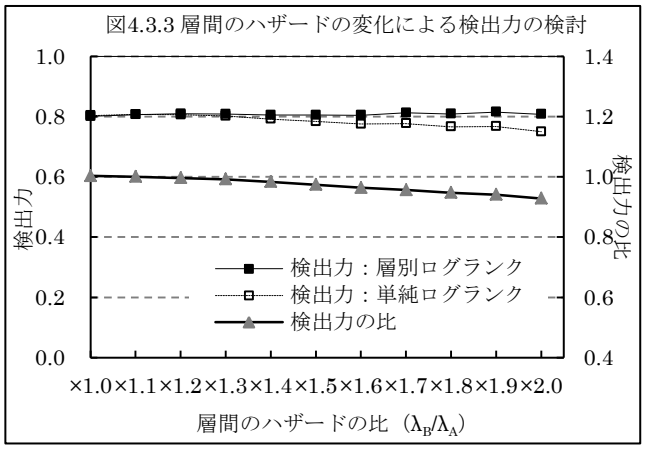
層 B のハザードを変化した時の層ごとの生存曲線について、1 層あたり 10 万症例としてシミュレーションした結果のグラフを、理論曲線に換えて図 4.3.1 および図 4.3.2 に示す。

【シミュレーション結果と検討】

結果を図 4.3.3 に示した。層間のハザードの比が増加するにつれて、単純ログランク検定の検出力は低下した。一方、層別ログランク検定では約 0.8 を維持した。このことより、層間でのハザードの比が大きくなればなるほど、層別ログランク検定を実施する意義が明確となることが示唆された。







しかしながら、層間のハザードの比が 1.3 倍までの検出力は拮抗しており、層間のハザードの違いがわずかであれば層別を行うことの意義は低いことも示唆された。

また、×1.0、×1.1 における検出力の比はそれぞれ 1.00337、1.00062 となり、ごくわずかではあるが層別を行う検定の検出力が低下した。

4.4.1 層間の例数がアンバランスな時の検出力の検討(層内ではバランス)

各群の層 A＋層 B の症例数を 150 から変化させず、層 A 層 B の症例数の割合を変化し、検出力の推移を確認した。ここでは、群 1 群 2 それぞれにおける層 A 層 B の症例数の割合は等しく保つ。

【データの説明】

	層 A	層 B
群 1	例数：N_{1A}＝下記の通り増加 ハザード：λ_{1A}=0.6	例数：N_{1B} ＝下記の通り減少 ハザード：λ_{1B}=0.9
群 2	例数：N_{2A} ＝下記の通り増加 ハザード：λ_{2A}=0.8	例数：N_{2B} ＝下記の通り減少 ハザード：λ_{2B}=1.26
観察期間：t=4 で打ち切り。 (N_{1A}:N_{1B},N_{2A}:N_{2B})：(5:145,5:145),(10:140,10:140),(25:125,25:125),(50:100,50:100), (75:75,75:75), (100:50,100:50), (125:25,125:25),(140:10,140:10),(145:5,145:5)		

【シミュレーション結果と検討】

結果を図 4.4.1.1 に示す。検出力はほぼ一定を保ったが、これはλ_{A}:λ_{2A}＝λ_{1B}:λ_{2B}と設定しているためと考えられる。例えばλ_{1A}:λ_{2A}＜λ_{1B}:λ_{2B}と設定していれば層 B の例数が多いほど検出力は向上するはずである。本件についても検討を行うため、λ_{2B}=1.4 とし、その他は先ほどの検証から変更しないシミュレーションを行

った。

シミュレーション結果を図 4.4.1.2 に示す。λ_{2B} を増加したことにより全体的に検出力は向上している。

λ_{2B}=1.4 とした場合、層 B の症例数が多い（=N_{1A} が少ない）ほど検出力が高く、層 B の症例数が少ないほど検出力が低くなる傾向となった。単純ログランク検定の検出力と層別ログランク検定の検出力の違いについては、全体的に層別ログランク検定の検出力が高い状態を維持したが、層の例数が均等に近いほど検出力の差は広がった。

図4.4.1.1 サンプル数がアンバランスな時の検出力

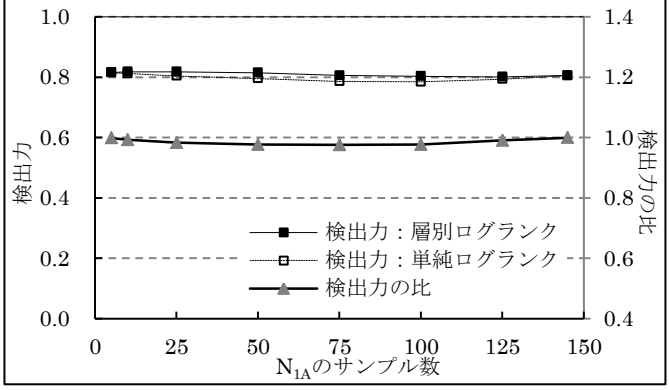
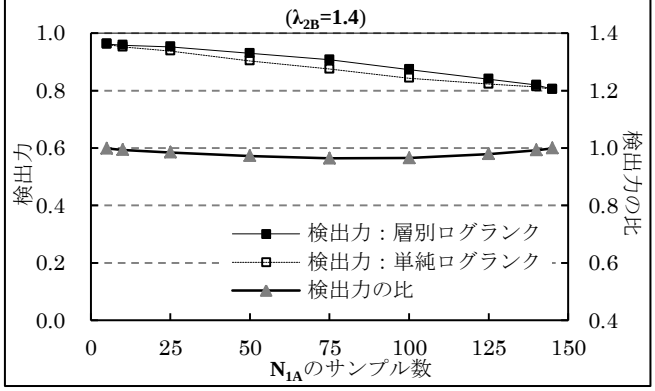


図4.4.1.2 サンプル数がアンバランスな時の検出力 (λ_{2B}=1.4)



4.4.2 層間の例数がアンバランスな時の検出力の検討(層内でアンバランス)

各群の層 A＋層 B の症例数を 150 から変化させず、層 A 層 B の症例数の割合を変化し、検出力の推移を確認した。ここでは、群 1 は層 A を徐々に増加、群 2 は層 B を徐々に増加させていく。なお、検出力の違いが大きく出たため、ここでは検出力の比の算出は行わなかった。

【データの説明】

	層 A	層 B
群 1	例数：N_{1A}＝下記の通り増加 ハザード：λ_{1A}=0.6	例数：N_{1B} ＝下記の通り減少 ハザード：λ_{1B}=0.9
群 2	例数：N_{2A} ＝下記の通り減少 ハザード：λ_{2A}=0.84	例数：N_{2B} ＝下記の通り増加 ハザード：λ_{2B}=1.26
観察期間：t=4 で打ち切り。 (N_{1A}:N_{1B},N_{2A}:N_{2B})：(5:145,145:5),(10:140,140:10),(25:125,125:25),(50:100,100:50), (75:75,75:75) , (100:50,50:100), (125:25,25:125),(140:10,10:140),(145:5,5:145)		

【シミュレーション結果と検討】

結果を図 4.4.2.1 に示す。単純ログランク検定では、N_{1A} および N_{2B} が増加するほど検出力が向上した。これは λ_{1A}:λ_{2B}＞λ_{2A}:λ_{1B} のため、λ_{A1} と λ_{2B} の症例数が増えるほど群間でのハザードの比が大きくなり、検出力が向上したためと思われる。対して層別ログランク検定では、層 A、層 B ごとに統計量を算出するため、例数がアンバランスな状態での統計量算出となってしまい、検出力低下を招いた[1]。

症例数のアンバランスが 1 層のみ発生する場合でも検討した。N_{1A}＝N_{1B}＝N_{2B}＝75 で固定し、N_{2A} のみ{5,10,25,50,75,125,140,145}と変化させた。シミュレーション結果を図 4.4.2.2 に示す。λ_{2A}＜λ_{2B} のため、単純ログランク検定では N_{2A} が少ないほどハザードに群間差が生じることとなり検出力が高い。逆に N_{2A} が多いほ

どハザードに群間差は小さくなるため検出力は減少した。

一方、層別ログランク検定では、層 A のアンバランスおよび例数過小のため、N_{2A} が少数の時に検出力が低下した。その後、N_{2A} の増加に合わせて検出力は向上した。

図4.4.2.1 層間の例数がアンバランスな時の検出力

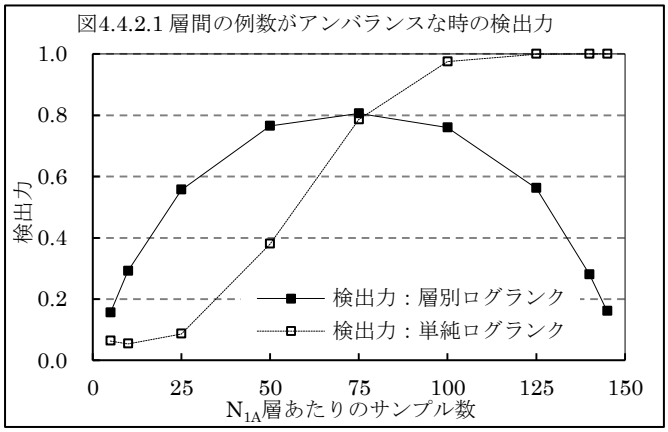
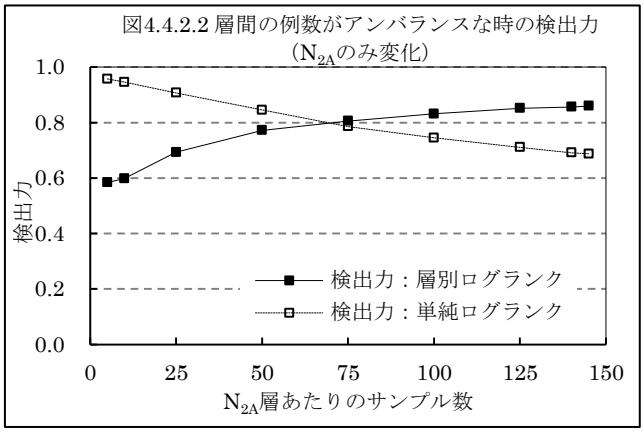


図4.4.2.2 層間の例数がアンバランスな時の検出力
（N_{2A}のみ変化）



4.5 層間の一様性が仮定できないときの検出力の検討

生存時間データを層別するにあたって、層間で治療効果が一様（同じ）であることが求められる[2]。ここでは層間の治療効果の一様性が仮定できないデータでの検出力について検討する。なお、治療効果が一様でないとは、層別因子と治療との間に交互作用があるということであり、この場合は交互作用について検討する必要がある。

層間の一様性を検定するプログラム以下の通り提示する[6]。本稿では、phreg プロシジャを用い、層間の交互作用について検定を行った。

SAS program4.5.1：phreg プロシジャによる層間の一様性を検定プログラム例

```
proc phreg data=入力データセット名 ;
  class 群別変数 層別変数 /ref=first param=ref;
  model 時間変数*打ち切り変数(打ち切り値) = 群別変数
        群別変数*層別変数 / rl type3(score) ;
  strata 層別変数 ;  run ;
```

【データの説明】

乱数データをそれぞれ 10 万通り作成、それぞれに層間の一様性の検定を実施し、有意（$p<0.05$）となったデータのみを対象として単純ログランク検定および層別ログランク検定を行い、検出力を算出・検討した。なお、層間の一様性の検定で優位になったということは、一様性があるとは言えないということである。

4.1 で示したデータと同様の乱数を使用して検討を行った。また、4.1 と同様、λ_{1B}= 1.2、λ_{2B}= 1.68 とした結果も示す。

10 万通りのデータ中、層間の一様性の検定で有意となったデータセット数を表 4.5 に示した。

【シミュレーション結果と検討】

結果を図 4.5.1 および図 4.5.2 に示す。全体的な推移は図 4.1.1 および図 4.1.2 と同様である。しかし、図 4.5.1 では 1 層あたりの例数が 50 症例未満の場合において、単純ログランク検定の検出力が層別ログランク検定の検出力を上回り、逆に層の例数が 50 症例以上の場合においては層別ログランク検定の検出力が単純ログラン

ク検定を上回っていた。図 4.5.2 では 25 症例以上にて層別ログランク検定の検出力が単純ログランク検定の検出力を上回っており、図 4.5.1 の場合と比べてより少ない症例数が閾値となった。このことより、層間の一様性の検定で有意となった場合も、層間のハザードの差が大きいほど、また症例数が多いほど層別ログランク検定の検出力が単純ログランクを上回ることが示唆された。なお、検出力が上回っていたとしても、交互作用に関する検討を放棄してよい理由にはならないと考えている。交互作用が認められた場合は、検出力とは全く別件として、交互作用に関する検討を行うべきと考える。

＜表 4.5 一様性の検定で有意となったデータセット数＞

層あたりの例数	有意になったデータセット数	
	λ_{1B}=0.9、λ_{2B}= 1.26	λ_{1B}= 1.2、λ_{2B}= 1.68
5	7,943	7,941
10	6,605	6,602
25	5,726	5,725
50	5,300	5,309
75	5,133	5,136

層あたりの例数	有意になったデータセット数	
	λ_{1B}=0.9、λ_{2B}= 1.26	λ_{1B}= 1.2、λ_{2B}= 1.68
100	5,079	5,089
125	5,203	5,208
150	5,082	5,089
200	4,961	4,988

図4.5.1 層間の一様性が仮定できないときの検出力
（症例数の変化）

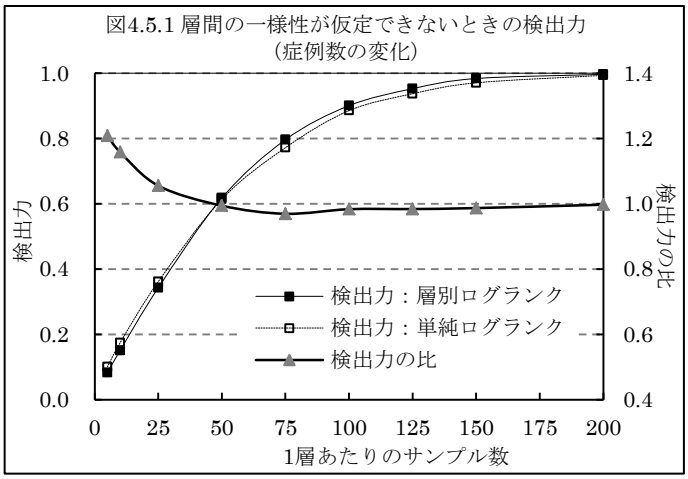
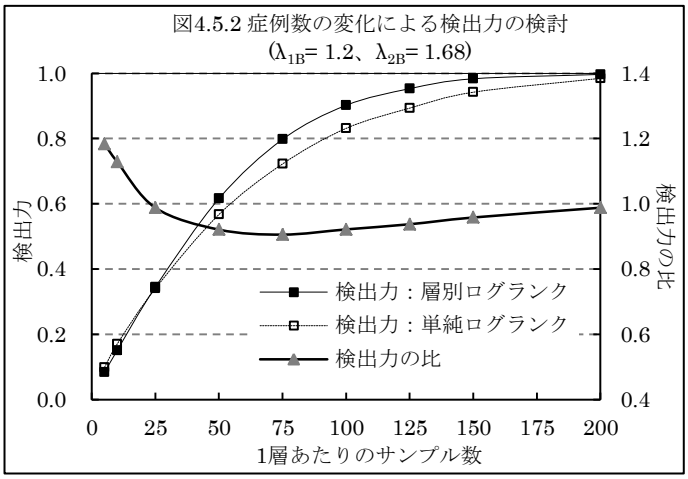


図4.5.2 症例数の変化による検出力の検討
(λ_{1B}= 1.2、λ_{2B}= 1.68)



5 まとめ

本稿ではシミュレーションにより、以下の内容が示唆された。

- 多くの場合、層間のハザードの差が大きいとき、層別ログランク検定を行うことにより検出力の低下を抑制する。
- 症例数のアンバランスが生じた場合や、層間の一様性が仮定できない場合において、層別ログランク検定を行うとかえって検出力が低下する場合がある。
- 層別ログランク検定を行うにあたっては、症例数が均一であることや、特に例数や層間のハザードの比が小さいときにおいては層間の一様性に留意する必要があることが示された。

最後に、本稿で示したシミュレーションデータが今後のシミュレーション実験や症例数設計の際に役立てば幸いである。

参考文献

[1] 赤澤宏平・柳川 堯 著．サバイバルデータの解析－生存時間とイベントヒストリデーター．近代科学社

[2] 大橋靖雄・浜田知久馬 著．生存時間解析 SAS による生物統計．東京大学出版会

[3] 監修/浜田知久馬,執筆/臨床評価研究会（ACE）基礎解析分科会．実用 SAS 生物統計ハンドブック[SAS8.2 および SAS9.1 対応]．サイエンティスト社

[4] 張方紅．SAS による生存時間解析の実務． SAS ユーザー総会 論文集 2012 167-184

[5] SAS Institute Inc． SAS/STAT(R) 9.2 User's Guide, Second Edition,The LIFETEST Procedure．
http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#lifetest_toc.htm

[6] 浜田知久馬・中西豊支・松岡伸篤．SASV9 の TPHREG を用いたメタアナリシス．SAS ユーザー総会 論文集 2004 165-191