生存時間分析についての研究

2008MI186 奥村和也
指導教員：松田眞一

1　はじめに

生存時間分析は、ある時点からイベントが起きるまでの時間を解析する分析方法である。研究分野として、工学分野においては機械システムや製品の故障などを、医学分野においては疾患の病気の再発や死亡などを対象としている。本研究の目的は、生存時間分析で用いられる３つのモデルについて理解し、実際の解析では乱数を用いてシミュレーションを行うことである。

2　生存時間分析の基本概念

2.1　生存関数

生存時間 T を累積確率分布関数 $F(t)$ と確率密度関数 $f(t)$ に従う非負の確率変数とする。イベントがある地点 t まで起きていない生存関数 $S(t)$ は

$$S(t) = 1 - F(t) = Pr(T > t) \tag{1}$$

と表わされる。(Armitage and Berry[1] 参照)

2.2　生存時間分析の分類

生存時間分析は、生存時間に影響を与える時間以外の共変量 (複数の要因、説明変数) がパラメータとして作成するモデルに導入されているかどうか、生存時間の分布形に特定の確率分布を仮定するかどうかによって、次の３つのモデルに分類することができる。

- ノンパラメトリックモデル (共変量を導入しない、分布を仮定しない)
- セミパラメトリックモデル (共変量を導入する、分布を仮定しない)
- パラメトリックモデル (共変量を導入する、分布を仮定する)

(Armitage and Berry[1] 参照)

2.3　生存関数の検定

２つの異なる集団の生存時間の観測値が得られたとき、群ごとの生存曲線の間の差の有意性について検定を必要とする場合がある。そのような時、最も広く用いられている検定方法はログ・ランク検定である。なお、ログ・ランク検定はノンパラメトリック検定であるため、ノンパラメトリックモデルに用いる。検定方法には、セミパラメトリックモデルは尤度比検定があり、パラメトリックモデルには（指数）分布の当てはめがある。

3　シミュレーション

3.1　手段

ある分布に従った乱数をそれぞれ２群ずつ発生させ、モデルごとの検定（ログ・ランク検定、尤度比検定、指数分布の当てはめ）を行い、２つの生存曲線に差はないという帰無仮説 (２つの生存曲線に差があるという対立仮説) が棄却されるかどうかを調べる。その作業を n 回繰り返し、群間で差が出る（p 値が 0.05 より小さくなる）回数とその確率を求めるプログラムを作り検証する。乱数には生存時間分析によく用いられる指数分布、ワイブル分布、対数正規分布の３つの分布を使用する。なお今回は有意水準５％で棄却されるかどうか調べるため、帰無仮説を想定した場合、p 値が 0.05 より大きくなれば、帰無仮説は棄却されず、群間でほとんど差がないということになる。対立仮説を想定した場合、p 値が 0.05 より小さくなれば、対立仮説は棄却されず、群間で差があるということになる。

3.2　準備

指数分布、ワイブル分布、対数正規分布を比較するため、平均と分散を揃える必要がある。

指数分布の平均と分散は次のようになる。

$$E(X) = \frac{1}{\lambda}, \qquad V(X) = \frac{1}{\lambda^2} \tag{2}$$

ワイブル分布の平均と分散は次のようになる。

$$\begin{aligned} E(X) &= a^{\frac{1}{b}}\Gamma\left(\frac{1}{b}+1\right), \\ V(X) &= a^{\frac{2}{b}}\left(\Gamma\left(\frac{2}{b}+1\right) - \Gamma^2\left(\frac{1}{b}+1\right)\right) \end{aligned} \tag{3}$$

対数正規分布の平均と分散は次のようになる。

$$E(X) = e^{\mu+\frac{\sigma^2}{2}}, \qquad V(X) = e^{2\mu+\sigma^2}(e^{\sigma^2}-1) \tag{4}$$

このように、指数分布は比率のパラメータ λ を、ワイブル分布は形状パラメータ b と尺度パラメータ a を、対数正規分布は正規分布の平均パラメータ μ と正規分布の標準偏差パラメータ σ を持っている。ここで、指数分布はパラメータが一つなので平均が決まれば分散もそれに応じて決まるが、ワイブル分布と対数正規分布はパラメータを二つずつ持っているので、平均と分散を揃えるようにパラメータの値を取る必要がある。例えば、指数分布の比率のパラメータ λ を 0.1 とすると、平均、分散はそれぞれ 10、100 となるので、ワイブル分布と対数正規分布も同じ平均、分散となるようにパラメータを設定しなければならないということである。平均 10、分散 100 なら、対数正規分布は $\mu = \log 10 - \frac{1}{2}\log 2$、$\sigma = \sqrt{\log 2}$ となり、ワイブル分布は a =8.3088、b =0.9505 となる。

3.3　帰無仮説の場合

分布に従った乱数を 50 個ずつ２群発生させ、繰り返しの回数を 1000 回として、プログラムを実行してみた。

帰無仮説:$\mu_1 = \mu_2$(２つの群の平均が同じ)

指数乱数、ワイブル乱数、対数正規乱数のどの乱数でも全体的に差が出ないということが分かった。ただし、指数乱数、ワイブル乱数はどのモデルでもそれほど差が出なかったのに対して、対数正規乱数はパラメトリックモデルと他の２つのモデルとで少し確率に差が出た。

3.4　対立仮説の場合

対立仮説:$\mu_1 \neq \mu_2$(２つの群の平均が異なる)
・ノンパラメトリックモデル
＜１群のパラメータが 0.10 の場合＞

表 1　３つの乱数での比較１ (横列: ２群のパラメータ)

	0.12	0.14	0.16	0.18	0.20
指数	0.141	0.392	0.638	0.819	0.928
ワイブル	0.082	0.18	0.387	0.526	0.711
対数正規	0.164	0.452	0.732	0.88	0.975

全体的にワイブル乱数は、指数乱数や対数正規乱数に比べると差が出にくいということが見て取れる。そして、指数乱数と対数正規乱数では、若干ではあるが対数正規乱数のほうが差が出るということが分かった。
＜１群のパラメータが 0.20 の場合＞

表 2　３つの乱数での比較２ (横列: ２群のパラメータ)

	0.23	0.26	0.29	0.32	0.35
指数	0.134	0.279	0.463	0.648	0.798
ワイブル	0.101	0.24	0.412	0.605	0.767
対数正規	0.104	0.277	0.524	0.739	0.874

ここでは、指数乱数、ワイブル乱数、対数正規乱数は最初どれも同じように確率が上昇していくということが見て取れる。しかしパラメータをずらしていくと、対数正規乱数の確率が他の２つの乱数と比べて急激に上昇するということが分かった。
・セミパラメトリックモデル
＜１群のパラメータが 0.10 の場合＞

表 3　３つの乱数での比較３ (横列: ２群のパラメータ)

	0.12	0.14	0.16	0.18	0.20
指数	0.14	0.388	0.632	0.815	0.928
ワイブル	0.081	0.177	0.381	0.522	0.707
対数正規	0.159	0.447	0.725	0.875	0.974

ノンパラメトリックモデルの時と同様に、全体的にワイブル乱数は、指数乱数や対数正規乱数に比べると差が出にくいということが見て取れる。そして、指数乱数と対数正規乱数では、若干ではあるが対数正規乱数のほうが差が出るということが分かった。
・パラメトリックモデル
＜１群のパラメータが 0.10 の場合＞
ノンパラメトリックモデル、セミパラメトリックモデルの時と同様に、全体的にワイブル乱数は、指数乱数や対数正規乱数に比べると差が出にくいということが見て取れる。しかし、指数乱数と対数正規乱数は、ノンパラメトリックモデル、セミパラメトリックモデルの時とは異なり、同じような確率の上昇をすることが分かった。

表 4　３つの乱数での比較４ (横列: ２群のパラメータ)

	0.12	0.14	0.16	0.18	0.20
指数	0.139	0.399	0.662	0.834	0.936
ワイブル	0.095	0.209	0.427	0.567	0.748
対数正規	0.124	0.378	0.655	0.833	0.943

＜１群のパラメータが 0.20 の場合＞

表 5　３つの乱数での比較５ (横列: ２群のパラメータ)

	0.23	0.26	0.29	0.32	0.35
指数	0.129	0.28	0.475	0.665	0.821
ワイブル	0.112	0.243	0.419	0.627	0.788
対数正規	0.093	0.233	0.458	0.669	0.82

ノンパラメトリックモデル、セミパラメトリックモデルの時とは異なり、ここでは、指数乱数、ワイブル乱数、対数正規乱数はどれも同じように確率が上昇していくということが分かった。

3.5　考察

パラメータをずらすと全体的に、対数正規乱数で最も差が出やすく、ワイブル乱数で最も差が出にくいという結果になった。ただし、ノンパラメトリックモデル、セミパラメトリックモデルは対数正規乱数の確率の上昇が大きいのに対して、パラメトリックモデルではパラメータを大きくしても他の２つの乱数とそれほど差は出なかった。これは、分布の当てはめに指数分布を用いたことが要因となっている。また、パラメトリックモデルでの結果は、ノンパラメトリックモデルやセミパラメトリックモデルの結果と比べて、確率に顕著な違いが見られた。これは、パラメトリックモデルが、ノンパラメトリックモデルやセミパラメトリックモデルと異なり、分布を仮定するモデルであることが一因となっていると考えられる。

4　おわりに

本研究を通して、医学統計の重要な解析方法である生存時間分析についてある程度理解できたと思う。解析を行う際は、実際のデータで解析を行いたかったが、やはり医療データなどを公表しているところは少ないため、乱数を用いることでデータの作成を行った。また、今回乱数には指数分布、ワイブル分布、対数正規分布を用いたが、その他にもガンマ分布や対数ロジスティック分布など生存時間モデルに使われる分布はまだあるので、今後比較、検討を行いたいと思う。

参考文献

[1] Armitage,P. and Berry,G. 著 (椿美智子・椿広計 訳): 医学研究のための統計的方法, サイエンティスト社, 2001