生存時間解析におけるコピュラに従うデータの分析の研究

M2018SS001 金武芽実
指導教員：松田眞一

1　はじめに

生存時間解析とは、ある基準からある目的までの反応が起きるまでの時間を解析対象としている統計的手法である．また，コピュラとは，各変数の依存関係を表現する関数である．このコピュラは生物統計において，相関性があるイベント時刻のモデリングにおいて用いられることがある．(Yan [7] 参照) コピュラの種類は，クレイトンコピュラ，ガンベルコピュラのように複数存在する．Emura and Chen [1] の研究では，統計ソフト R を用いてクレイトンコピュラに従う実データの生存時間解析を行っている．しかし，クレイトンコピュラ以外のコピュラに従うデータの生存時間解析や，シミュレーションは行っていない．本研究では，R を用いて，コピュラに従うデータに対してログランク検定のシミュレーションを行い，結果の相違が見られるのか比較，検証を行う．

2　先行研究

Emura and Chen [1] において，R の compoumd.Cox パッケージ内の肺がんの患者 125 人の生存データの分析を行っている．フォローアップ期間中，38 人の患者は死亡し，残りの 87 人の患者は治験終了を含む打ち切りとなった．また，125 人のデータのうち，63 人を traning set とし，クレイトンコピュラに基づく単変量の Cox 回帰を適用した．このとき，推定したクレイトンコピュラのパラメータは $\hat{\alpha} = 18$ (4 章での θ と同じ)，Kendall's $\tau = 0.90$ となった．分析の結果，97 の遺伝子のうち，p 値を基準に 16 の遺伝子の選択を行った．

3　生存時間解析

生存時間解析とは，ある時刻からの疾患の再発，または死亡といった目的の反応が起きるまでの時間について解析を行う統計的手法である．ここでの時間とは，ある個人のあるイベントが起こるまでの年，月，週，または日数を意味する．また，死亡，発症，疾患の再発などをイベント (failure) とする．(Kleinbaum and Klein [4] 参照)

3.1　打ち切り

打ち切りは個人の生存時間について，以下のような場合に起こるとされている．

1. 試験終了時，その人にイベントが発生しない場合
2. 試験期間中，その人がフォローアップ不能の場合
3. 死亡や死亡原因が興味のあるイベントでない，あるいは薬の副作用などにより，その人が試験から脱落する場合

3.2　ログランク検定

ログランク検定はカプランマイヤー (KM) 曲線の全般的な比較の判断指標となる統計量を用いた χ^2 検定である．2 群のログランク検定における検定統計量は式 (1) となる．

$$\frac{(O_i - E_i)^2}{Var(O_i - E_i)} \quad (i = 1, 2) \tag{1}$$

O_i：第 i 群の観測度数
E_i：第 i 群の期待度数
$i = 1, 2$：群番号

4　コピュラ

n 個の確率変数 $X_1, \ldots, X_n$ について，周辺分布関数をそれぞれ $F(x_1), \ldots, F(x_n)$ とし，同時分布関数を $F(x_1, \ldots, x_n)$ とする．このとき，以下の関係がある．

スクラーの定理 (Sklar's theorm)
周辺分布関数 $F_1, \cdots, F_n$ をもつ連続な n 変量分布関数について，以下の関係を満たす関数 C が一意に存在する．

$$\begin{aligned} &Pr(X \leq x_1, \cdots, X \leq x_n) \\ &= F(x_1, \cdots, x_n) = C(F_1(x_1), \cdots, F_n(x_n)) \end{aligned} \tag{2}$$

この関数 C がコピュラである．ここで，任意の $u_i = F_i(x_i)$ $(u_i \in [0, 1], i = 1, \cdots, n)$ について

$$C(u_1, \cdots, u_n) = F(F_1^{-1}(u_1), \cdots, F_n^{-1}(u_n)) \tag{3}$$

と与えられる．このとき関数 C は，各周辺分布か区間 [0,1] の一様分布となるような同時分布関数である．(戸坂・吉羽 [6] 参照)

4.1　コピュラの種類

コピュラには主に 2 つの種類がある．(矢田・浜田 [8] 参照)

1. 相関行列で変量間の依存構造を表現するコピュラ
2. 1 種類のパラメータを用いて変量間の依存構造を表現するコピュラ

1 の代表例は正規コピュラなどがあり，2 の代表例は，アルキメディアンコピュラがある．アルキメディアンコピュラとは，生成素 ϕ を用いて式 (4) のように表現され，クレイトンコピュラ，ガンベルコピュラなどがある．

$$C(u_1, \cdots, u_n) = \phi^{-1}(\phi(u_1) + \cdots + \phi(u_n)) \tag{4}$$

2 つのコピュラのパラメータ θ と生成素の関係は表 1 のようにまとめられる．(Emura and Chen [1] 参照) これらのコピュラはそれぞれ式 (5)，式 (6) のように定義される．

クレイトンコピュラ

$$C(u_1, \cdots, u_n) = \left(\sum_{i=1}^{n} u_i^{-\theta} - n + 1 \right)^{-\frac{1}{\theta}} \tag{5}$$

表 1 パラメータの範囲と生成素

	範囲	生成素
クレイトン	$\theta > 0$	$\frac{t^{-\theta}-1}{\theta}$
ガンベル	$\theta \geq 0$	$\{-\log(t)\}^{\theta+1}$

ガンベルコピュラ

$$C(u_1, \cdots, u_n) = \exp\left[-((\log(u_1))^\theta + \cdots + (\log(u_n))^\theta)^{\frac{1}{\theta}}\right] \tag{6}$$

クレイトンコピュラとガンベルコピュラに従う乱数の散布図を図 1 と図 2 に示す．Kendall's τ（2 節参照）を 0.75 と設定し，乱数を 1000 個生成させた．

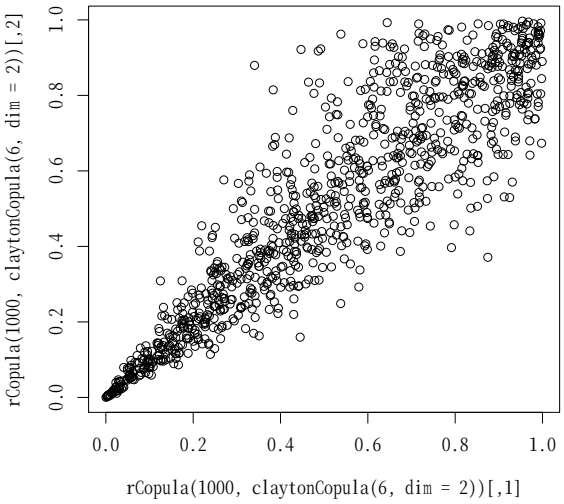

図 1 クレイトンコピュラの散布図

図 1 より、クレイントンコピュラは左下が強く、右上で弱い．図 2 より，ガンベルコピュラは右上で強く、左下で弱い.（戸坂・吉羽 [6] 参照）

4.2 順位相関

順位相関とは，各変量のデータの値そのものではなく，ある基準による各変量のデータの順位に基づく相関のことであり，その一つに Kendall's τ がある．アルキメデアンコピュラの Kendall's τ は表 2 のように簡略化される．(Emura and Chen [1] , Haurd *et al.* [3] 参照)

表 2 アルキメデアンコピュラの Kendall's τ

	Kendall's τ
クレイトン	$\frac{\theta}{\theta+2}$
ガンベル	$1-\theta^{-1}$

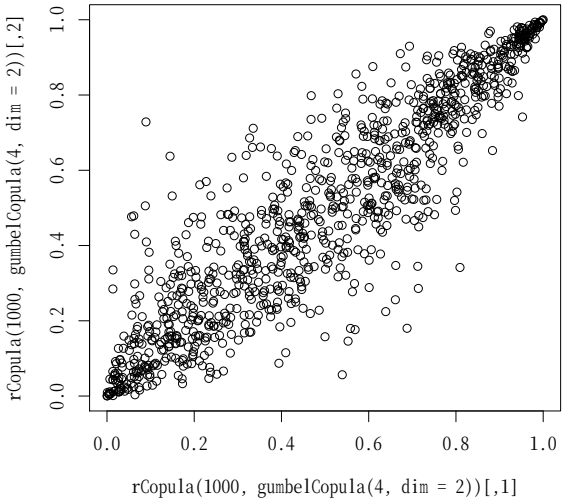

図 2 ガンベルコピュラの散布図

4.3 乱数の生成

R 内でのアルキメデアンコピュラの乱数の生成方法はマーシャル=オルキン法を用いている.（Hofert *et al.* [2] 参照）本研究では copula パッケージを用いて生成した．

5 プログラミング

今回，ログランク検定をするために survival パッケージを，コピュラに従う乱数を発生させる copula パッケージを使用した．本研究のシミュレーションの流れは以下の通り．

1. コピュラに従う乱数の生成
2. 発生させた乱数をパラメータ 1 の指数分布に変換
3. 各群のイベント発生時間，打ち切り時間を生成
4. 打ち切りの発生
5. ログランク検定用のデータを生成
6. ログランク検定を行い，p 値を求める

各群のイベント発生時間を生成するために指数分布のパラメータ λ を使用し，$\lambda = 1.5$ を初期設定としている．一方，打ち切り時間の指数分布のパラメータ η は脱落率が所定の値になるように調整する．この η を脱落率パラメータと呼ぶことにする．イベント発生時間，打ち切り時間は，指数分布に変換した後にパラメータ λ と η で割ることによって生成する．

5.1 打ち切りの実現

本研究での打ち切りの実現方法を示す．イベント発生時間と打ち切り発生時間を比較し，個人ごとに最小値を求める．この求めた最小値がイベント発生時間と等しいならばイベントが発生，等しくないならば打ち切りとする．また治験は 2 年以上行われないと設定しているので，最小値が 2 以上の場合は 2 で打ち切りとする．脱落率とは，最小値が 2 未満で打ち切りとなった人数の割合を指

す．クレイトンコピュラ，ガンベルコピュラの脱落率パラメータをそれぞれ表 3，表 4 に示す．この脱落率パラメータは，シミュレーションを 10000 回実行し，脱落率の平均を求めて探索した．なお，0%については十分小さい数値であれば達成できるため，統一して与えた．

表 3 クレイトンコピュラの脱落率パラメータ

	0%	10%	20%
τ=0.75	0.00001	0.858	1.069
τ=0.5	0.00001	0.5	0.743
τ=0.3	0.00001	0.3172	0.548
τ=0.25	0.00001	0.282	0.509
τ=0.1	0.00001	0.197	0.413

表 4 ガンベルコピュラの脱落率パラメータ

	0%	10%	20%
$\tau = 0.75$	0.00001	0.7934	1.0583
$\tau = 0.5$	0.00001	0.4393	0.739
$\tau = 0.3$	0.00001	0.285	0.5562
$\tau = 0.25$	0.00001	0.2566	0.5184
$\tau = 0.1$	0.00001	0.1928	0.4168

6 シミュレーション

乱数を生成する際の Kendall's τ は 0.75，0.5，0.3，0.25，0.1 とし，脱落率は各群ごとに 0%，10%，20%とした 9 通りの組み合わせで実行した．村井ら [5] などによると，実際の治験での脱落率は大きくても約 20%であったため，最大の脱落率を 20%と設定した．またシミュレーション回数は 100000 回とした．

7 クレイトンコピュラに従う場合の結果

7.1 2 群の生存時間に差がない場合

各群 50 人，100 人，150 人，200 人，250 人としシミュレーションを実行した．結果を Kendall's τ ごとに表 5〜表 7 に示す．

表 5 Kendall's $\tau = 0.75$ の有意確率

脱落率	50 人	100 人	150 人	200 人	250 人
(0%,0%)	0.0539	0.0517	0.0516	0.0511	0.0501
(0%,10%)	0.0575	0.0601	0.0640	0.0669	0.0711
(0%,20%)	0.0836	0.1150	0.1496	0.1812	0.2166
(10%,0%)	0.0568	0.0590	0.0644	0.0684	0.0721
(10%,10%)	0.0529	0.0516	0.0512	0.0510	0.0504
(10%,20%)	0.0653	0.0773	0.0913	0.1041	0.1171
(20%,0%)	0.0834	0.1152	0.1493	0.1828	0.2191
(20%,10%)	0.0644	0.0774	0.0910	0.1056	0.1186
(20%,20%)	0.0525	0.0516	0.0504	0.0510	0.0500

表 5〜表 7 より，脱落率が第 1 群と第 2 群で等しい場合，Kendall's τ の値によらず有意水準を保つことが分かった．

表 6 Kendall's $\tau = 0.5$ の有意確率

	50 人	100 人	150 人	200 人	250 人
(0%,0%)	0.0539	0.0517	0.0516	0.0511	0.0501
(0%,10%)	0.0561	0.0573	0.0603	0.0618	0.0648
(0%,20%)	0.0685	0.0825	0.1005	0.1168	0.1336
(10%,0%)	0.0556	0.0564	0.0604	0.0634	0.0658
(10%,10%)	0.0527	0.0517	0.0512	0.0511	0.0503
(10%,20%)	0.0573	0.0622	0.0676	0.0729	0.0773
(20%,0%)	0.0680	0.0830	0.1005	0.1170	0.1338
(20%,10%)	0.0574	0.0624	0.0676	0.0733	0.0676
(20%,20%)	0.0525	0.0518	0.0505	0.0515	0.0503

表 7 Kendall's $\tau = 0.1$ の有意確率

	50 人	100 人	150 人	200 人	250 人
(0%,0%)	0.0532	0.0517	0.0516	0.0511	0.0501
(0%,10%)	0.0535	0.0525	0.0523	0.0525	0.0526
(0%,20%)	0.0539	0.0540	0.0550	0.0551	0.0558
(10%,0%)	0.0532	0.0521	0.0522	0.0530	0.0524
(10%,10%)	0.0530	0.0516	0.0509	0.0516	0.0501
(10%,20%)	0.0528	0.0521	0.0518	0.0518	0.0513
(20%,0%)	0.0529	0.0535	0.0550	0.0555	0.0560
(20%,10%)	0.0523	0.0526	0.0519	0.0516	0.0519
(20%,20%)	0.0525	0.0521	0.0518	0.0508	0.0503

しかし，脱落率の差が 20% ある場合は，Kendall's τ が高いと有意水準が守られてないことが分かった．加えて人数も多くなるとその傾向がかなり強く見られる．

7.2 2 群の生存時間に差がある場合

第 2 群について，生存時間が第 1 群よりも良くなる場合，悪化する場合についてシミュレーションを行った．第 1 群のイベント発生時間のパラメータを 1.5 とし，良くなる場合はパラメータを低くし，悪化する場合は高く設定した．このパラメータは 1000 回シミュレーションを行ったときに検出力が 10%，50%，90%に近くなる値としている．Kendall's τ が 0.75 のときの変更した第 2 群のパラメータとそのパラメータごとの脱落率パラメータを表 8，結果を表 9 に示す．なお，表 3 などと同様に，0%については十分小さい数値で達成できるため統一して与えた．

表 8 脱落率パラメータ (τ=0.75)

	0.7	1	1.3	1.7	2.3	2.9
0%	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
10%	0.3616	0.5435	0.7336	0.9835	1.35541	1.709
20%	0.45536	0.6835	0.9153	1.2188	1.66256	2.097

表 9 より，脱落率の差が 20%の場合について，第 2 群の方が良くなっているときは第 2 群の脱落率が 20%のときの方が検出されやすい．しかし，第 2 群の方が悪化している場合は，第 2 群の脱落率が 20%のときの方が検出されにくくなっていることが分かる．

表 9 $\tau = 0.75$ の検出力

	変更した第 2 群のパラメータ λ_2					
	0.7	1	1.3	1.7	2.3	2.9
(0%,0%)	0.9400	0.4874	0.1099	0.0972	0.5489	0.8933
(0%,10%)	0.9473	0.5324	0.1392	0.0706	0.4342	0.8089
(0%,20%)	0.9642	0.6308	0.2150	0.0535	0.2874	0.6669
(10%,0%)	0.8988	0.3973	0.0814	0.1260	0.6080	0.9164
(10%,10%)	0.9094	0.4402	0.1025	0.0904	0.4911	0.8425
(10%,20%)	0.9348	0.5403	0.1616	0.0580	0.3375	0.7125
(20%,0%)	0.7987	0.2612	0.0552	0.2006	0.7148	0.9520
(20%,10%)	0.8159	0.2966	0.0622	0.1487	0.6075	0.9022
(20%,20%)	0.8581	0.3879	0.0944	0.0859	0.4493	0.7984

7.3 人数が異なる場合

Kendall's τ を 0.75，0.1，第 1 群を 50 人，第 2 群 250 人とした場合のシミュレーション結果を表 10 に示す．

表 10 第 1 群 50 人第 2 群 250 人の有意確率

	$\tau = 0.75$	$\tau = 0.1$
(0%,0%)	0.0525	0.0525
(0%,10%)	0.0618	0.0545
(0%,20%)	0.1173	0.0563
(10%,0%)	0.0567	0.0513
(10%,10%)	0.0518	0.0518
(10%,20%)	0.0815	0.0530
(20%,0%)	0.0949	0.0504
(20%,10%)	0.0668	0.0505
(20%,20%)	0.0516	0.0513

表 10 より，Kendall's τ が 0.75 で，脱落率の差が 20%の場合，第 2 群の脱落率が 20%のときは有意確率が 10%を超えるが，第 1 群の脱落率が 20%のときは 10%越えないという結果になった．

8 ガンベルコピュラに従う場合

ガンベルコピュラに従い，人数を各群 50 人としたときの結果を表 11 に示す．

表 11 50 人のときのガンベルコピュラの有意確率

	τ				
	0.75	0.5	0.3	0.25	0.1
(0%,0%)	0.0536	0.0539	0.0532	0.0534	0.0536
(0%,10%)	0.0763	0.0653	0.0577	0.0564	0.0535
(0%,20%)	0.1496	0.1039	0.0732	0.0674	0.0550
(10%,0%)	0.0778	0.0656	0.0608	0.0563	0.0536
(10%,10%)	0.0537	0.0530	0.0529	0.0530	0.0532
(10%,20%)	0.0774	0.0661	0.0552	0.0572	0.0538
(20%,0%)	0.1518	0.1042	0.0715	0.0661	0.0549
(20%,10%)	0.0768	0.0652	0.0570	0.0559	0.0533
(20%,20%)	0.0539	0.0533	0.0525	0.0527	0.0530

表 11 より，脱落率の差があると Kendall's τ の値が高いほど，クレイトンコピュラより有意水準を守りにくい．しかし，2 群の脱落率が等しいときはクレイトンコピュラと同様に τ の値に関係なく有意水準は守られやすい．

9 考察

ガンベルコピュラは人数が少なくても，Kendall's τ が高い，脱落率に差がある場合は有意水準を守りにくいことが分かった．

クレイトンコピュラは人数が少ないならば，脱落率の差が 20%であっても，有意確率は 10%となっているが，人数が多くなる場合は守られにくくなる．また，人数が異なる，2 群間での生存時間に差がある場合，脱落率によって結果に差があるので，どちらが脱落率が高いのか考えて分析する必要があることが分かった．

Emura and Chen [1] では，数学的な単純さからクレイトンコピュラを推奨しているが，今回の結果からログランク検定においてもガンベルコピュラは有意水準を守りにくい．ガンベルコピュラに従うデータである場合，より有意確率が高くなることに注意して分析する必要があることが分かった．また，どちらの場合でも，人数や Kendall's τ の値，脱落率について気を付けなければならないと考えられる．

10 まとめ

本研究では，クレイトンコピュラの方が優れているということが分かった．しかし，他のコピュラとの比較や，実際の脱落率が設定した脱落率と異なっている場合について検証する必要があると考えられる．

参考文献

[1] Emura, T., Chen, Y.: Analysis of Survival Data with Dependent Censoring –Copula-Based Approaches–, Springer, 2018.

[2] Hofert, M., Kojadnovic, I., Maechler, M., Yan, J. and Neslehova., J.: Multivariate Dependent with Copulas, CRAN, 2019.

[3] Huard, D., Evin, G. and Favre, A.: Bayesian copula selection, *Computational Statistics & Data Analysis*, **51**, 809-822, 2006.

[4] Kleinbaum, D.G. and Klein, M. 神田英一郎・藤井朋子訳：エモリー大学クラインバウム教授の生存時間解析，サイエンティスト社，2015.

[5] 村井正大・牧野勲嗣 他：歯肉炎および辺縁性歯周炎に対するパスター剤「アセス A」の二重盲検法による薬効評価,『日本歯周病学会会誌』**24**(3)，490-515，1982.

[6] 戸坂凡展・吉羽要直：コピュラの金融実務での具体的な活用方法の解説,『金融研究』，115-160，2005.

[7] Yan, J.: Enjoy the Joy of Copulas: With a Package copula, *Journal of Statistical Software*, **21**(4), DOI:10.18637/jss.v021.i04, 2007.

[8] 矢田真城･浜田知久馬:SAS を用いたコピュラに従う擬似乱数の生成,『SAS ユーザー総会アカデミア/テクノロジー&ソリューションセッション論文集』，643-656，2014.