深夜アニメ視聴データに対する生存時間分析の適用

○道仙光輝　小山友介（芝浦工業大学）

Application of Survival Analysis to Late-Night Anime Viewing Data

* K. Dosen and Y. Koyama (University of Shibaura)

概要－　本研究では，深夜アニメの視聴実態の調査を目的として，深夜アニメ視聴データに対して生存時間分析を実行した．生存曲線の作成には Kaplan-Meier 曲線を利用する．得られた生存曲線にクラスター分析を実行することで，アニメの視聴実態と性別や年齢などの属性にどのような関係があるかを調査した．調査の結果，アニメには 4 つのクラスターが存在し，クラスターごとに属性から受ける影響に差があることが判明した．

キーワード:　アニメ，生存時間分析，クラスター分析

1. はじめに

日本のアニメーション(以下，アニメ)作品は国内外から高い評価を受けている．2016年に公開された「君の名は」は原作が存在しないオリジナル作品であるのにもかかわらず，250億円の興行収入[1)]を獲得した．近年では，2020年10月に公開された「鬼滅の刃-無限列車編-」が2020年12月現在で324億円[1)]の興行収入を獲得したことが記憶に新しい．この「鬼滅の刃」ブームは普段アニメを見ることがないような層をも巻き込んでいることが特徴であり，過去では「進撃の巨人」が類似のブームを引き起こしている．これらのブームは，深夜枠でのアニメ放送がきっかけであると考えられている．[2)]深夜枠のメインの客層であるマニア層を超え，一般視聴者層を取り込むブームとなる可能性を秘めていることから，深夜枠のアニメ(以下，深夜アニメ)は業界内外で高い注目を受けている分野である．

しかし，深夜アニメの観られ方については経験則などの情報が多く，不明瞭な状態である．深夜アニメに関する経験則の例として，「3話切り」という言葉がある．「3話切り」とは，アニメの第3話を視聴した時点で，それ以降の視聴を止めてしまう状態を指す言葉である．しかし，2017年には，アニメの第1話を視聴した時点で視聴を止めてしまう「1話切り」が増えているという報告も上がっている．[3)] これは，年間の作品数の増加や生活体系の変化などで，視聴者が一本のアニメに対して割くことができる時間が少なくなったことが原因だと考えられている．

また，深夜アニメを取り巻く環境にも変化が生じている．20104年には国内でAmazon PrimeやNetflixなどの配信サービスが急速に普及した．[2)]さらに録画機能を備えた液晶テレビの発売により，録画視聴の敷居が下がり深夜アニメを視聴する方法は過去に比べて格段に増えた．

このように，深夜アニメの観られ方には不明瞭な点が多く，視聴形態も変化していることから，アニメの観られ方に注目を向けた分析が必要である．そこで，本研究では深夜アニメの「観られ方」の分析を目的として，アニメ視聴データを用いた生存時間分析を行う．分析結果はKaplan-Meier生存曲線を利用して解釈を行い，それらにWard法を用いた階層クラスタリングを実行することで，アニメ作品の分類を行っていく．

2. 生存時間分析[4)]

アニメ作品の視聴データを分析するにあたって，本研究では生存時間分析(survival analysis)を活用する．生存時間分析とは，目的変数を，”あるイベントが起きるまでの時間”としてデータを解析する一連の統計手法である．時間とは，分析対象を観察し始めてからイベントが発生するまでの期間であり，年，月，週などや，イベント発生時の状態を意味する．イベントとは，死亡や疾病の発症，サービスの離脱や機器の故障など，ある個人に発生しうる任意の事象である．同期間中に複数のイベントを想定することも可能ではあるが，本研究では一つのイベントのみに注目する．生存時間分析では，時間変数を生存時間と呼び，イベントをfailureと呼ぶ．また，観察期間中にイベントが発生しない場合や，外的要因によって観察が中断されるなど，個人の生存時間についての明確な情報を取得することができなかった状態を打ち切り(censoring)と呼ぶ．一般的に，イベントの発生または打ち切りの情報は0，1などに変換され，二値変数dで表現する．

本研究においては，アニメ作品の継続視聴話数を生存時間として扱い，継続視聴の断念をイベントとして扱う．アニメ作品の視聴においての打ち切りは，視聴を断念することがなかったという状態である．

2.1. 生存関数とハザード関数

生存時間分析には，生存関数(survivor function，$S(t)$)とハザード関数(hazard function，$h(t)$)という２つの関数を利用する．生存関数$S(t)$とは，観察対象がある特定期間tよりも長く生存する確率である．確率変数Tが特定の時間tを超える確率と言い換えることもできる．ハザード関数$h(t)$とは，観察対象が特定期間tよりも長く生存しているという条件のもと，単位時間あたりにイベントが発生する瞬間的な可能性である．生存関数$S(t)$とハザード関数$h(t)$は対応関係にあり，以下の式で表すことができる．

$$S(t) = \left[-\int_0^t h(u)du\right]$$

$$h(t) = -\left[\frac{d\,S(t)/dt}{S(t)}\right]$$

2.2. Kaplan-Meier 生存曲線

生存時間データのプロット，解釈には生存確率推定値を利用する．この値を利用してプロットされたデータを Kaplan-Meier 生存曲線(以下，KM 曲線)と呼ぶ．生存確率推定値$\hat{S}(t_{(f)})$は積極限式によって求められる他，一つ前の failure 時間$(t_{(f-1)})$での生存確率に対して検討している failure 時間を超えて生存する場合の条件付き確率$\hat{P}r$を掛けることで計算することもできる．

$$\hat{S}(t_{(f)}) = \prod_{i=1}^{f} \hat{P}r[T > t_{(i)} \mid T \geq t_{(i)}]$$

$$= \hat{S}(t_{(f-1)}) \times \hat{P}r(T > t_{(f)} \mid T \geq t_{(f)})$$

生存曲線は，時間$t = 0$の$S_{(t)} = 1$から減少していき，$t = \infty$では$S_{(t)} = 0$となる．イベントが発生しない人がいる場合は観察終了時に 0 に到達しない．実際のデータでは，生存曲線のデータはステップ関数で表現される．

2.3. Log-rank 検定

複数の KM 曲線に対して，統計的な差が存在するかを検証する手法として，Log-rank 検定を利用する．Log-rank 検定の検定統計量は他のχ^2検定と同様に，カテゴリー毎の観察度数(O)と期待度数(E)の差の合計である．3 群以上の比較も可能ではあるが，本研究では 2 群間の比較にのみ，Log-rank 検定を利用する．2 群間を比較する場合，Log-rank 検定統計量を求めるために，2 群のうちどちらかの「観察度数−期待度数」の値をすべての failure 時間について合計したもの$(O_i - E_i)$を利用する．Log-rank 検定統計量を求める式を以下に示す．

$$Log\text{-}rank\ statistic = \frac{(O_i - E_i)^2}{Var(O_i - E_i)}$$

Log-rank 検定の帰無仮説(H_0)は，「2 つの生存曲線には全体として差がない$(S_1(t) = S_2(t))$」である．それに対し，対立仮説(H_1)は「2 つの生存曲線には全体として差がある$(S_1(t) > S_2(t))$」である．求められた p 値があらかじめ定めた基準以下であった場合，帰無仮説は棄却される．生存曲線に対する検定法は他にも Wilcoxon や Tarone-Ware，Peto，Flemington-Harrington などがある．Log-rank 検定はそれらの検定法に対し，後期に発生した failure を重く評価するという特徴がある．なお，本研究では Log-rank 検定のみを使用する．

3. データの紹介

本研究では，東芝映像ソリューション株式会社から提供された，液晶テレビ REGZA の視聴データを利用する．このデータは，2019 年 1 月～6 月の 6 カ月間に放送されたアニメ作品 86 作品の視聴データを格納している．提供されたデータは 3 種あり，それぞれ(A)番組視聴実績データ，(B)機器属性データ，(C)番組メタデータである．

(A)番組視聴実績データには，リアルタイム視聴のフラグである flag_live と，録画視聴のフラグである flag_rec などの作品ごとの視聴情報が格納されている．視聴の判定は 10 分以上視聴または再生することが条件となっており，リアルタイム視聴が 5 分，録画視聴が 5 分の場合は視聴したと判定されない．また，今回の分析では，リアルタイム視聴･録画視聴を区別せず，1 話から視聴し続けた話数のみを生存時間として扱う．

(B)機器属性データには，REGZA を利用しているユーザーの性別や年齢，設定地域や画面サイズの情報が格納されている．

(C)番組メタデータには，放送された番組の放送開始時期や放送開始時間，番組タイトルやシリーズ名の情報が格納されている．

また，本データはビデオリサーチ社が収集している視聴率とはデータ収集方法が異なる．視聴率は，日本の世代構造を可能な限り忠実に反映した母集団で収集されたデータであり，本データはネットワーク対応 REGZA をネットワークに接続し，視聴データの取得に許諾した家庭が母集団となる．[5)]

3.1. 成形後データ

2019 年 1 月から 6 月に放送された 86 作品の視聴実績データを Python にて成形した結果を以下に示す．

Table 1: 成形後のデータ

列名	詳細
mem_l	機器(ユーザー)ID
pro_l	作品に振り分けられた番号
lif_l	作品の視聴話数
eve_l	イベント発生の有無(0,1)
gen_l	機器に登録された性別
age_l	機器に登録された年齢
wat_l	総合視聴作品数

性別は男性，女性を 0，1 のダミー変数に変換したものを利用する．年齢は 31 歳未満を若年層として扱い，31 歳以上 46 歳未満を中年層，46 歳以上を高齢層とし，それぞれ 0~2 までの数値に変換した．分析対象機器数は 55043 台であり，関東での集計機器全体の 32%が分析対象となった．

4. 結果

2019 年 1 月から 6 月の間に放送された作品のうち，深夜に放送された 83 作品のアニメに生存時間分析を実行し，Kaplan-Meier 曲線を作成した．全体の傾向として，一話での視聴者の離脱が最も大きく，その後は緩やかに離脱が発生する作品が大半を占めた．全作品の生存率と離脱率の平均値を以下に示す．離脱率は視聴継続中の視聴者のうち，イベント(離脱)が発生した割合である．

Table 2: 生存曲線の平均

	ep1	ep3	ep6	ep9	ep12
生存率	51.1%	32.9%	24.5%	19.4%	17.5%
離脱率	48.9%	16.1%	8.4%	8.2%	1.7%

ほとんどの作品は 1 話放送終了時点で半分の視聴者が離脱しており，最終話での視聴者は 2 割にも満たないことがわかる．離脱率は 4 話以降 10%弱で安定しており，10 話や 12 話での離脱率は 2%以下と非常に小さい値となった．しかし，9 話と 11 話の離脱率は他に比べて高い．これは，9 話，11 話において生存率が著しく下落した作品があり，それに引きずられたと考えられる．

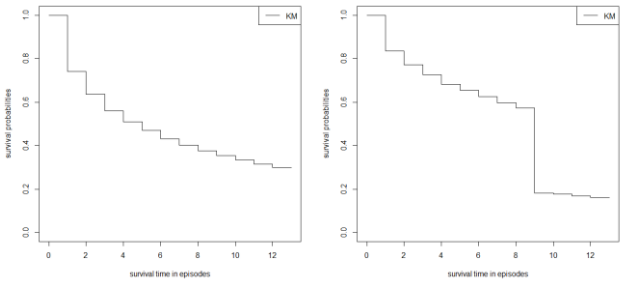

Fig 1: 平均的な生存曲線と特異な生存曲線

Fig1(右)の様に断層のような生存曲線が発生した原因として，放送時間の変更が挙げられる．特異な生存曲線が描かれた作品は，通常 25:35 に放送が開始されるが，9 話のみ放送時間が変更され，25:59 に放送が開始された．これにより，REGZA の機能である全録(1 話から 12 話まで一気に録画予約する機能)が正しく機能せず，録画視聴が不可能になり，視聴者の大幅な離脱が発生したと考えられる．

4.1. ユーザー属性を交えた分析

ユーザー属性のうち，性別，年齢，視聴作品数を用いて分析を行っていく．性別，年齢には無回答やその他などのデータも含まれているが，それらの属性を持つデータは除外して分析を行った．

4.1.1. 性別

成形データ内の gen_l 列を利用し，性別で層化した Kaplan-Meier 曲線を作成した．基本的にはユーザー属性を考慮しない Kaplan-Meier 曲線に近い結果を示していたが，男女で生存曲線の優劣がハッキリとしている作品が確認できた．女性の生存曲線が優れた結果を残していた作品と，男性の生存曲線が優れた結果を残していた作品それぞれを以下に示す．

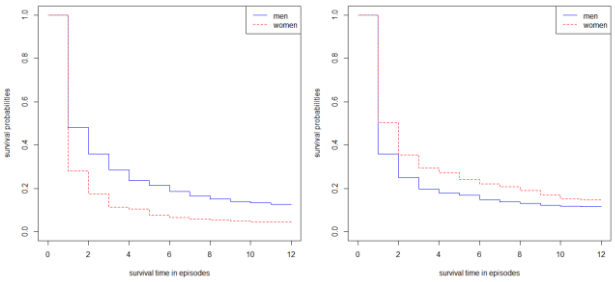

Fig 2: 男性向け作品(左)と女性向け作品(右)

このような生存曲線が描かれる場合，男性向けや女性向けと明確にターゲッティングを行っている作品が多い．例えば，男性アイドルが活躍するアニメでは女性の生存曲線が男性に比べ優秀な結果を残している．

次に，アニメ視聴に対する意欲が，性別で異なっているかを確認するため，平均生存率の比較を行う．男性，女性それぞれの平均生存率を以下に示す．

Table 3: 性別ごとの生存曲線平均

	ep1	ep3	ep6	ep9	ep12
男性	52.2%	33.9%	25.4%	20.2%	18.0%
女性	45.8%	28.2%	20.7%	16.0%	13.8%

男性に比べ，女性の平均生存率が全体的に 5%程低い結果となり，男性の方がアニメ視聴に対する意欲が高いと考えられる．これは，1 つのクールにおいて放送される女性向け作品の本数が，男性向け作品に比べて半分以下であることが原因であると考えられる．また，女性の生存曲線が男性の生存曲線に比べ優れている作品は，元々の生存率が低い傾向にある．

4.1.2. 年齢

成形データ内の age_l 列を利用し，年齢で層化した Kaplan-Meier 曲線を作成した．性別で層化した Kaplan-Meier 曲線と変わらず，ユーザー属性を考慮しない生存曲線に近い結果を示していた．性別に比べ顕著な差がある作品は少なく，ほとんどの作品で年齢による生存曲線のバラつきが確認できなかった．

性別と同様に，年齢ごとにアニメ視聴に対する意欲に差があるかを確認するため，平均生存率の比較を行う．年齢ごとの平均生存率を以下に示す．

Table 4: 年齢ごとの生存曲線平均

	Ep1	ep3	ep6	ep9	ep12
若年層	49.0%	30.6%	22.0%	16.9%	14.6%
中年層	51.7%	33.4%	25.1%	20.1%	18.0%
高齢層	51.9%	34.0%	25.9%	20.8%	18.4%

高齢層，中年層の差は 1%以内であるのに対し，若年層の平均生存率が 3%~4%程低い結果となった．また，中年層は 11 話~12 話にかけて視聴離脱が発生していないことが確認できる．以上のことから，アニメ視聴に対する意識は，高齢層と中年層がほぼ同じであり，若年層がやや離脱しやすいという結果となった．

4.1.3. 視聴作品数

データ G 内の wat_l 列を利用し，クール間の視聴数で層化した Kaplan-Meier 曲線を作成した．前述した 2 つの属性とは異なり，単話視聴(クール間で 1 作品のみ視聴)の Kaplan-Meier 曲線は独特な形状をしている作品が多く確認できた．他の属性と同じ Kaplan-Meier 曲線を描く作品も少数存在しているが，単作品視聴の生存曲線に比べ，クール間で複数作品を視聴しているユーザーの生存曲線の方が優れていることが多かった．

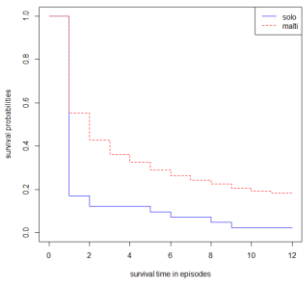

Fig 5: 視聴作品数で層化した生存曲線

アニメ視聴に対する意欲の差を確認するため，平均生存率の比較を行う．単作品視聴，複数話視聴それぞれの平均生存率を以下に示す．

Table 5: 視聴作品数ごとの生存曲線平均

	ep1	ep3	ep6	ep9	ep12
単作品	27.5%	15.4%	11.7%	8.7%	7.7%
複数作品	53.8%	34.9%	26.1%	20.8%	18.5%

単作品視聴の平均生存率が圧倒的に低い．その原因として，一話の大量離脱が挙げられる．1 話において，他の生存曲線では平均 50%近くが離脱しているのに対し，単作品視聴の離脱率は 72%と非常に高い数値となっている．それに対し，複数作品視聴の生存率は，全体の生存率に比べ高いことがわかる．以上のことから，アニメ視聴に対する意識は，複数作品視聴の方が高く，単作品視聴は圧倒的に低いという結果となった．

4.2. Log-rank 検定

各属性において，複数の生存曲線に統計的に有意な差があるかを確認するため，Log-rank 検定を実行した．検定の結果，83 作品中 79 作品に差があるという結果が出力された．しかし，Kaplan-Meier 曲線を確認したところ，視覚的には明確に差がないと確認できる作品においても，統計的には差があるという結果になってしまっていた．この現象は，サンプル数が多すぎることが原因であると考えられる．本来，Log-rank 検定は 60~100 のサンプル数で行うものであり，サンプル数が 2000~4000 ほどあるデータの場合，実際には意味がないわずかなさでも検定に通ってしまったためである．

5. 考察

アニメ視聴データに対して生存時間分析を実行した結果，視聴離脱が最も発生しやすい期間は 1 話から 3 話の間であり，瞬間的な離脱発生率が最も高いのは 1 話放送終了後であった．この離脱が発生しやすい期間が 3 話切りという経験則を生んだ要因であると考えられる．生存時間分析の観点では，1 話切りと 3 話切りどちらも肯定するような結果が得られた．この 1 話切りと 3 話切りは名前が似通っているが，根拠としているものは異なる．1 話切りが根拠とするものは，KM 曲線でも明らかにされた「1 話での大量離脱がほぼ全ての作品で発生している」というものである．それに対し，3 話切りの根拠として確認できた分析結果は，「1～3 話の期間が視聴者の離脱ピークであり，その後の離脱は非常に緩やかである」というものであった．これらの情報から，1 話切りと 3 話切りの関係性は以下の様であると予測できる．

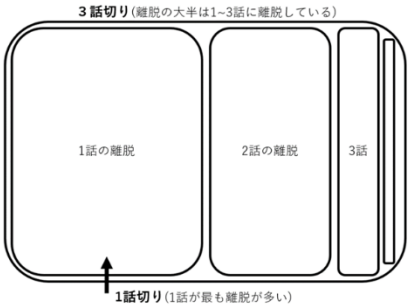

Fig 6: 1 話切りと 3 話切りの関係性

図 12 から，3 話切りは 1 話切りの根拠を含んだ包括的な視聴傾向であり，1 話切りは 3 話切りの中でも最も結果が明白なものを根拠としている視聴傾向であることが予測できる．1 話切り，3 話切りは共存可能であり，根拠は同じものであるが，観点が異なっているだけであると考えられる．

6. 今後の課題と結論

本研究では，深夜アニメの視聴実態調査を目的として，アニメ視聴データに対し生存時間分析を実行した．また，性別，年齢，視聴作品数など属性別の生存曲線の作成も同時に行った．その結果，視聴データの観点でアニメ化が成功したと考えられる作品は，全体の 20%程であることが判明した．今後の課題として，利用データの拡張と Cox 比例ハザードモデルへの適用が挙げられる．

今回の研究では，分析対象を 1 話から継続して視聴しているユーザーのみに生存時間分析を行った．しかし，それでは生存時間分析の利点を十分に活かしているとは言い難い．例えば，2 話から視聴開始したユーザーを分析対象に加えるといったように途中開始データを利用することで，分析対象数を増やすことができる．また，生存時間分析には Cox 比例ハザードモデルという非常に頑健なモデルが存在している．本研究においても，全作品に Cox 比例ハザードモデルの適用を試みたが，作品ごとに比例ハザード性を満たす変数が異なることが多く，作品間の比較が困難になってしまったため，適用を断念した．層別 Cox や他のモデルとの併用などで，Cox 比例ハザードモデルが適用できれば，さらに詳細な分析が可能になると考えられる．

参考文献

1) http://www.kogyotsushin.com/archives/alltime/
2) 数土直志：デジタルコンテンツ白書 2020，66/71，一般社団法人デジタルコンテンツ協会
3) http://m.timeon.jp/analytics/
4) D. G. Kleinbaum et al ,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
5) https://www.videor.co.jp/press/2020/200206.html