2025 年 1 月 29 日

タイタニック号生存に関する解析

泉 七海

目次

1. 序論 3

1-1. 解析の背景 3

1-2. 解析の目的 3

1-3. 解析の方針 3

2. 使用するデータ 4

3. 解析結果と考察 5

3-1. 要約統計量 5

3-2. 度数表 8

3-3. ロジスティック回帰 10

4. 結論 12

1. 序論

1-1. 解析の背景

1912 年 4 月に沈没したタイタニック号の悲劇は、史上最も有名な海難事故の一つである。この事件では多数の犠牲者を出し、社会的・歴史的な関心を集め続けている。特に、乗客の客室クラス、性別、年齢といった個人属性が生存にどのように影響を与えたかについては、長年にわたり多くの分析が行われてきた。本解析では、これらの要因を再検証し、生存者の特徴を統計的に分析する。

1-2. 解析の目的

本解析の目的は、タイタニック号の乗客データを用いて生存者の特徴を統計的に分析し、生存率に影響を与えた要因を明らかにすることである。具体的には、要約統計量、度数表、ロジスティック回帰分析を用いた解析を行い、性別、客室クラス、年齢、運賃、乗船地といった要因が生存確率に及ぼす影響を検証する。

1-3. 解析の方針

本解析では、タイタニック号の乗客データを統計的に分析するために、要約統計量、度数表、ロジスティック回帰分析の 3 つの手法を用いる。まず、要約統計量を算出することで、データの基本的な傾向を把握し、生存者と非生存者の特徴を比較する。次に、度数表を用いて性別や客室クラスなどのカテゴリ変数ごとの生存率を可視化し、要因ごとの影響を直感的に示す。最後に、ロジスティック回帰分析を用いることで、生存率に影響を与えた要因を定量的に評価し、それぞれの変数が生存確率に及ぼす影響の大きさを検証する。これらの手法を組み合わせることで、生存率を左右した要因を包括的に分析することを目指す。

2. 使用するデータ

本解析で使用するデータは、タイタニック号の乗客情報をまとめたものである。このデータには、891 人の乗客の個人情報（年齢、性別、客室クラスなど）と、生存の有無が含まれている。

データに含まれている変数は以下の 12 個である。

PassengerId (乗客 ID)

Survived (生死 0:死亡, 1:生存)

Pclass (客室クラス 1:1st, 2:2nd, 3:3rd)

Name (氏名)

Gender (性別)

Age (年齢)

SibSp (同乗している兄弟,配偶者の数)

Parch (同乗している親,子供の数)

Ticket (チケット番号)

Fare (運賃)

Cabin (客室番号)

Embarked (出港地)

なお、データには欠損値が存在する。特に、Age で 177 件、Cabin で 687 件、Embarked で 2 件の欠損が確認されている。本解析では、欠損値のあるデータを削除して解析を行う。

3. 解析結果と考察

3-1. 要約統計量

```
title "生存と性別と客室クラスにおける年齢と運賃の要約統計";
proc means data=titanic mean median var std min max;
    class Survived Gender Pclass;
    var Age Fare;
run;
```

生存と性別と客室クラスにおける年齢と運賃の要約統計

MEANS プロシジャ

Survived	Gender	Pclass	Obs 数	変数	平均	中央値	分散	標準偏差	最小値	最大値
0	female	1	3	Age	25.6666667	25.0000000	576.3333333	24.0069434	2.0000000	50.0000000
				Fare	110.6041667	151.5500000	5029.68	70.9202637	28.7125000	151.5500000
		2	6	Age	36.0000000	32.5000000	166.8000000	12.9151074	24.0000000	57.0000000
				Fare	18.2500000	17.0000000	48.5750000	6.9695767	10.5000000	26.0000000
		3	72	Age	23.8181818	22.0000000	164.6978114	12.8334645	2.0000000	48.0000000
				Fare	19.7730931	14.4791500	212.3511931	14.5722748	6.7500000	69.5500000
	male	1	77	Age	44.5819672	45.5000000	209.0265027	14.4577489	18.0000000	71.0000000
				Fare	62.8949104	42.4000000	3606.31	60.0525449	0	263.0000000
		2	91	Age	33.3690476	30.5000000	147.8199943	12.1581246	16.0000000	70.0000000
				Fare	19.4889648	13.0000000	247.1356350	15.7205482	0	73.5000000
		3	300	Age	27.2558140	25.0000000	147.2753749	12.1357066	1.0000000	74.0000000
				Fare	12.2044693	7.8958000	120.6681187	10.9849041	0	69.5500000
1	female	1	91	Age	34.9390244	35.0000000	174.8480879	13.2230136	14.0000000	63.0000000
				Fare	105.9781593	82.1708000	5585.90	74.7388969	25.9292000	512.3292000
		2	70	Age	28.0808824	28.0000000	162.9373903	12.7646931	2.0000000	55.0000000
				Fare	22.2889886	23.0000000	124.1204771	11.1409370	10.5000000	65.0000000
		3	72	Age	19.3297872	19.0000000	151.3698543	12.3032457	0.7500000	63.0000000
				Fare	12.4645264	9.4687500	35.8167990	5.9847138	7.2250000	31.3875000
	male	1	45	Age	36.2480000	36.0000000	223.1063138	14.9367437	0.9200000	80.0000000
				Fare	74.6373200	35.5000000	10219.58	101.0919479	26.2875000	512.3292000
		2	17	Age	16.0220000	3.0000000	382.0899600	19.5471215	0.6700000	62.0000000
				Fare	21.0951000	18.7500000	96.6788797	9.8325419	10.5000000	39.0000000
		3	47	Age	22.2742105	25.0000000	133.5361926	11.5557861	0.4200000	45.0000000
				Fare	15.5796957	8.0500000	232.0256353	15.2323877	0	56.4958000

本表は、生存、性別、客室クラスごとの乗客の年齢および運賃に関する要約統計量を示している。この結果から、1等客室の乗客は他のクラスよりも生存率が高く、特に女性の生存率が顕著に高いことが確認された。一方で、3等客室の乗客は生存率が低く、特に男性の死亡率が高いことが分かった。また、運賃が高いほど生存率が高い傾向が見られ、経済的な地位が救助の際に影響を与えた可能性が示唆された。

```
title "生死ごとの年齢と運賃の要約統計量";
proc means data=titanic mean median var std min max;
    class Survived;
    var Age Fare;
run;
```

生死ごとの年齢と運賃の要約統計量

MEANS プロシジャ

Survived	Obs 数	変数	平均	中央値	分散	標準偏差	最小値	最大値
0	549	Age	30.6261792	28.0000000	200.8486984	14.1721099	1.0000000	74.0000000
		Fare	22.1178869	10.5000000	985.2195092	31.3882065	0	263.0000000
1	342	Age	28.3436897	28.0000000	223.5309652	14.9509520	0.4200000	80.0000000
		Fare	48.3954076	26.0000000	4435.16	66.5969981	0	512.3292000

生存者と非生存者の年齢と運賃を比較した結果、生存者の平均年齢は28.34歳、非生存者は30.63歳であり、わずかに非生存者の方が高い傾向が見られた。運賃については、生存者の平均が48.40であるのに対し、非生存者の平均は22.12と低く、生存者の方がより高額のチケットを購入していたことが示唆された。

```
title "生死と性別ごとの年齢と運賃の要約統計量";
proc means data=titanic mean median var std min max;
    class Survived Gender;
    var Age Fare;
run;
```

生死と性別ごとの年齢と運賃の要約統計量

MEANS プロシジャ

Survived	Gender	Obs 数	変数	平均	中央値	分散	標準偏差	最小値	最大値
0	female	81	Age	25.0468750	24.5000000	185.4660218	13.6185910	2.0000000	57.0000000
			Fare	23.0243852	15.2458000	616.0963131	24.8212875	6.7500000	151.5500000
	male	468	Age	31.6180556	29.0000000	197.5716787	14.0560193	1.0000000	74.0000000
			Fare	21.9609929	9.4166500	1050.40	32.4097992	0	263.0000000
1	female	233	Age	28.8477157	28.0000000	200.9326861	14.1750727	0.7500000	63.0000000
			Fare	51.9385734	26.0000000	4109.10	64.1022561	7.2250000	512.3292000
	male	109	Age	27.2760215	28.0000000	272.4085220	16.5048030	0.4200000	80.0000000
			Fare	40.8214844	26.2875000	5091.67	71.3559670	0	512.3292000

性別ごとに生存者と非生存者の年齢および運賃を分析した結果、女性の生存率が男性よりも著しく高いことが明らかとなった。特に1等および2等客室の女性の生存率は高く、3等客室の男性の生存率が極めて低いことが分かった。また、女性の運賃の平均値は男性よりも高く、上級クラスの女性が救助の優先度が高かったことを示唆している。

3-2. 度数表

```
title "性別ごとの生存確率度数表";
proc freq data=titanic;
    tables Survived*Gender / norow nopercent;
run;
```

性別ごとの生存確率度数表

FREQ プロシジャ

度数 列のパーセント	表：Survived * Gender		
	Gender		
Survived	female	male	合計
0	81 25.80	468 81.11	549
1	233 74.20	109 18.89	342
合計	314	577	891

性別ごとの生存確率を確認した結果、女性の生存率は74.2%であったのに対し、男性の生存率は18.89%と著しく低かった。この結果は、救助活動の際に「女性と子供が優先された」という方針が統計的に裏付けられることを示している。

```
title "性別ごとの客室クラス分布度数表";
proc freq data=titanic;
    tables Pclass*Gender / norow nopercent;
    run;
```

性別ごとの客室クラス分布度数表

FREQ プロシジャ

度数 列のパーセント	表：Pclass * Gender		
	Gender		
Pclass	female	male	合計
1	94 29.94	122 21.14	216
2	76 24.20	108 18.72	184
3	144 45.86	347 60.14	491
合計	314	577	891

客室クラスごとの男女の分布を確認すると、1 等および 2 等客室では女性の割合が比較的高い一方、3 等客室では男性の割合が大幅に高いことが分かった。これは、3 等客室には移民層の乗客が多く含まれていたことを示唆している。

3-3. ロジスティック回帰

```
proc logistic data=titanic;
    class Gender(ref='male') Pclass(ref='3') Embarked(ref='S');
    model Survived(event='1') = Pclass Gender Age Fare Embarked;
run;
```

LOGISTIC プロシジャ

モデルの情報	
データセット	WORK.TITANIC
応答変数	Survived
応答の水準数	2
モデル	binary logit
最適化の手法	Fisher's scoring

読み込んだオブザベーション数	891
使用されたオブザベーション数	712

反応プロファイル		
順番	Survived	度数の合計
1	0	424
2	1	288

分類変数の水準の情報			
分類	値	デザイン変数	
Gender	female	1	
	male	−1	
Pclass	1	1	0
	2	0	1
	3	−1	−1
Embarked	C	1	0
	Q	0	1
	S	−1	−1

モデル収束状態
収束基準(GCONV=1E-8)は満たされました。

モデルの適合度統計量		
基準	切片のみ	切片と共変量
AIC	962.904	658.674
SC	967.472	695.218
−2 Log L	960.904	642.674

包括的帰無仮説: BETA=0 の検定			
検定	カイ 2 乗値	自由度	Pr > ChiSq
尤度比	318.2302	7	<.0001
スコア	278.9918	7	<.0001
Wald	190.5832	7	<.0001

効果に対する Type 3 分析

効果	自由度	Wald カイ 2 乗値	Pr > ChiSq
Pclass	2	57.8085	<.0001
Gender	1	143.3593	<.0001
Age	1	21.5252	<.0001
Fare	1	0.0032	0.9546
Embarked	2	3.9563	0.1383

オッズ比の推定

効果	点推定	95% Wald 信頼限界	
Pclass 1 vs 3	11.232	5.849	21.570
Pclass 2 vs 3	3.548	2.183	5.769
Gender female vs male	12.390	8.206	18.707
Age	0.965	0.950	0.979
Fare	1.000	0.995	1.004
Embarked C vs S	1.642	0.968	2.786
Embarked Q vs S	0.726	0.260	2.024

最尤推定値の分析

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗値	Pr > ChiSq
Intercept		1	1.1655	0.3283	12.6039	0.0004
Pclass	1	1	1.1903	0.2014	34.9286	<.0001
Pclass	2	1	0.0381	0.1553	0.0602	0.8062
Gender	female	1	1.2584	0.1051	143.3593	<.0001
Age		1	−0.0361	0.00779	21.5252	<.0001
Fare		1	−0.00013	0.00229	0.0032	0.9546
Embarked	C	1	0.4375	0.2405	3.3098	0.0689
Embarked	Q	1	−0.3791	0.3533	1.1509	0.2834

予測確率と観測データの応答との関連性

一致の割合	85.4	Somers の D	0.709
不一致の割合	14.5	ガンマ	0.709
タイの割合	0.0	Tau-a	0.342
組	122112	c	0.854

ロジスティック回帰分析の結果、客室クラスと性別が生存確率に最も大きな影響を与えていたことが確認された。1等客室の乗客の生存確率は3等客室よりも11.23倍高く、女性の生存確率は男性の12.39倍高いことが示された。また、年齢が1歳増加するごとに生存確率は0.965倍低下し、高齢者が避難に不利であった可能性が示唆された。一方で、運賃や乗船地の影響は統計的に有意ではなかった。モデルの予測一致率は85.4%であり、生存確率を比較的高精度で予測できることが示唆された。

4. 結論

本解析では、タイタニック号の乗客データを用いて生存確率に影響を与えた要因を分析した。その結果、客室クラスと性別が生存率に最も大きな影響を与える要因であることが明らかになった。特に 1 等客室に乗船した女性の生存率が高く、3 等客室の男性の生存率が極めて低かったことが統計的に示された。また、年齢も有意な影響を持つが、客室クラスや性別ほどの影響は確認されなかった。一方で、運賃や乗船地の影響は統計的に有意ではなかった。これらの結果は、タイタニック号の救助活動において「女性と子供の優先」および「客室クラスによる救助格差」が存在していたことを示唆している。今後の解析では、交互作用の分析や、さらに詳細なデータを用いた解析を行うことで、より精緻な知見を得ることが期待される。