割引総報酬関数の分布

船　木　洋　一

不確かな現実を確率モデルでモデル化しても、そのモデルから得られる期待値以外の情報を用いないで利用されていることが多い。マルコフ決定過程のモデルでも、単独で、あるいは種々の条件と組み合わせて、期待値を基準とした最適政策の選択が行われている。一方、単純に期待値のみで政策を選ぶことが望ましいとはかぎらないことはよく知られている。本稿ではマルコフ決定過程のモデルの割引総報酬関数の分布の様子を見ることによってそのことを調べた。

クリスタルボールというシミュレーションソフトがあり、確率的な変動が組み合わされて起こる結果を分析するのに広く用いられている。それは、クリスタルボールが変動する結果の分布をシミュレーションにより与えてくれるからである。クリスタルボールのその特色を利用して、本稿では割引総報酬関数の分布を求め、それに考察を加えた。

1．オモチャ屋の例

はじめにハワードが使用したオモチャ屋の例を見てみよう（参考文献［1］）。データは表1.1である。この例は最適政策による割引総報酬関数の分布に特徴がある。

表1．1

状態	戦略	推移確率		直接報酬		直接期待報酬
		状態1	状態2	状態1	状態2	
状態1	戦略1（広告する）	0.5	0.5	9.0	3.0	6.0
（成功したオモチャ）	戦略2（広告しない）	0.8	0.2	4.0	4.0	4.0
状態2	戦略1（広告する）	0.4	0.6	3.0	−7.0	−3.0
（不成功のオモチャ）	戦略2（広告しない）	0.7	0.3	1.0	−19.0	−5.0

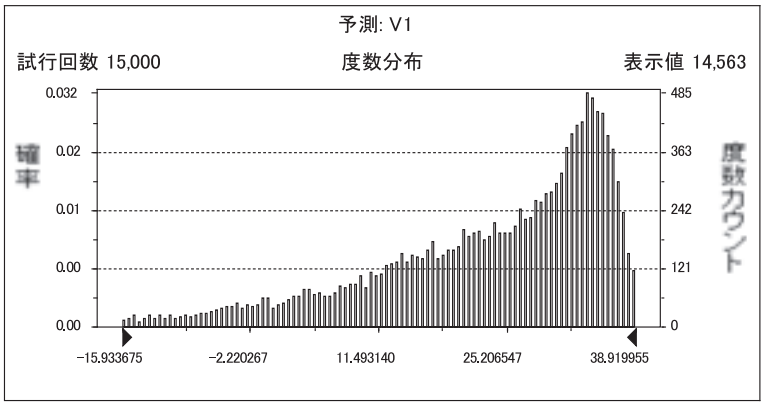

図１．２

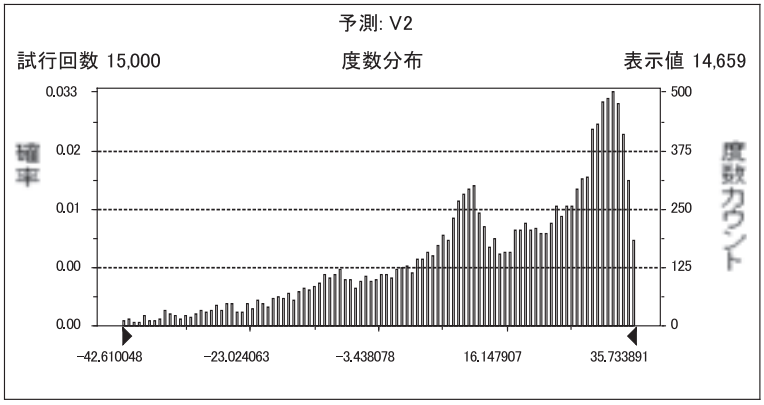

図１．３

ディスカウントファクターが0.9の時は、状態１、状態２でともに戦略２を用いるのが最適である。図１.２、図１.３はそれを用いたときの、それぞれ状態１、状態２の総報酬関数の分布の図である。これらの図から、ピークが値の高い方によっていることがわかる。一方、値の低い方にも長い裾ができているが、総報酬関数の値が０以下になる確率も、この政策の時が一番少ないことが計算により求められる。総報酬が多い方が良いという問題では大変好都合の政策である。最適政策はディスカウントファクターに依存して決まるが、ディスカウントファクターが0.5の時は、状態１、状態２でともに戦略１を用いるのが最適である。最適政策による総報酬関数は図１.４、図１.５のようになる。0.9の時の最適政策による分布とは違い、状態２からスタートした場合には値の低い方に分布のピークがある。状態だけを観察して戦略を決める定常政策の中から最適な政策を選択すると仮定すると、このとき総報酬関数の値が０以下になる確率が最も小さくなるのは、状態１から出発したときには状態１で戦略１を、状態２では戦略２を用いる政策であり、また状態２から出発したときには状態１、２ともに戦略２を用いる政策である。この例から示唆されるように総報酬関数の分布を調べることにより、状況にあったよりよい政策の選択ができるように思われる。

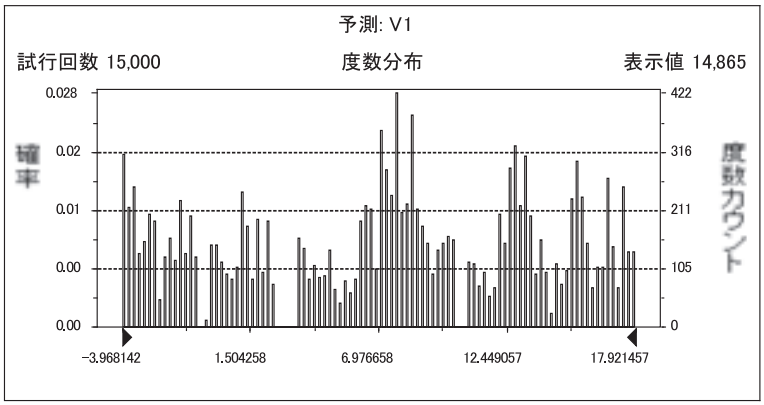

図１．４

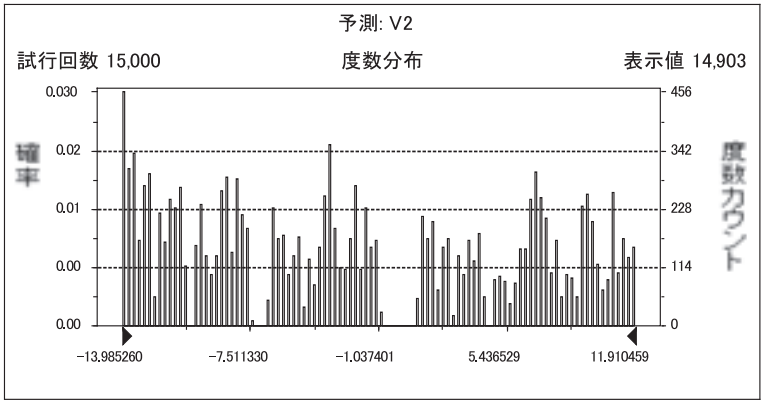

図１．５

２．ディスカウントファクターによる分布の違い

次に、割引総報酬関数はディスカウントファクターに依存するので、その値を変えて総報酬関数の分布を調べてみよう。ここでのデータは表２.１である。ハワードのオモチャ屋の例は２状態の例であったが、３状態で調べてみよう。割引総報酬の期待値は各ディスカウントファクターに対して表２.２のようになる。カイ二乗検定を基準として、どのタイプの連続分布が当てはまるかを調べた結果が表２.４、表２.５である。これらは最大試行回数10000回、ランダムシード999、ラテンハイパーキューブでサンプルサイズ500としたときの結果である。Ｐ値の大きい方から３候補載せてある。当てはまりの良さなどシミュレーションの条件によって変動するようである。シミュレーションソフトでは当てはめの様子をグラフで見ることができる(図２.３)。

表をみると、ベータ分布が近似の候補として各ディスカウントファクターのところで登場している。ここでのベータ分布は密度関数が

$$f(x)=\frac{\Gamma(\alpha,\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{1}{s}\left(\frac{x}{s}\right)^{\alpha-1}\left(1-\frac{x}{s}\right)^{\beta-1}\qquad \alpha>0,\ \beta>0,\ s>x>0$$

で表される分布である。この分布は二つのパラメータαとβが等しいときには左右対称になり、相対的にαが大きいときには左に裾を引き、βが大きいときには右に裾を引き、さらにα、βが大きければ大きいほど尖度が大きくなることが知られている。本節の例では、ベータ分布との当てはまり具合と、そのベータ分布の特徴から、ディスカウントファクターが大きくなるに従って、尖度が高くなるような傾向にあることがわかる。試行回数の違いでP値が変動することや推移確率の違いなどにより分布が違ってくることを考慮すると残念ながらより一般的なことは言えないようである。

表２．１

推移確率				直接期待報酬
	状態1	状態2	状態3	
状態1	0.3	0.3	0.4	10
状態2	0.5	0.0	0.5	20
状態3	0.6	0.3	0.1	100

表２．２

割引期待総報酬				
	ディスカウントファクター			
	0.7	0.8	0.9	0.95
状態1	111.552	181.4256	389.9374	806.1681
状態2	124.1195	194.1727	402.8334	819.1282
状態3	185.9322	254.0062	460.8035	876.2071

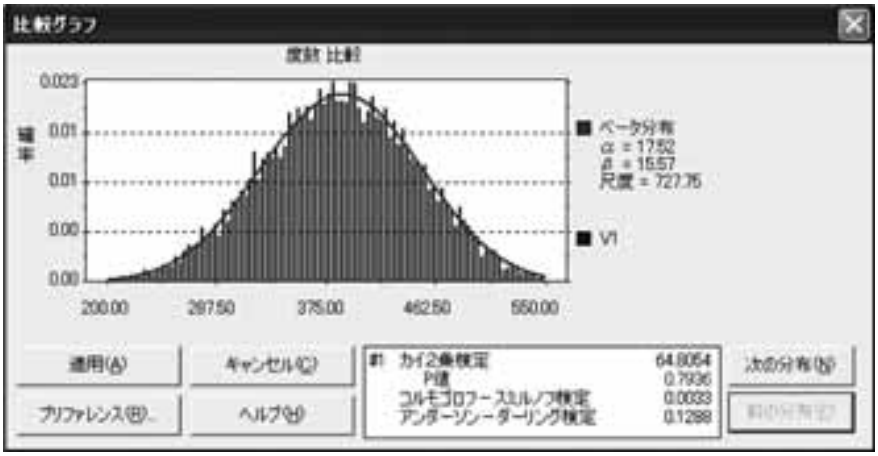

図２．３

表２．４

	ﾃﾞｨｽｶｳﾝﾄﾌｧｸﾀｰ 0.7				ﾃﾞｨｽｶｳﾝﾄﾌｧｸﾀｰ 0.8			
		パラメータ他	χ二乗検定	P値		パラメータ他	χ二乗検定	P値
V1	ベータ分布	α=5.65 β=5.60 尺度=222.30	305.3604	0.0000	ベータ分布	α=8.77 β=8.37 尺度=354.65	79.8312	0.3297
	ワイブル分布	位置=33.91 尺度=87.41 形状=2.63	418.4726	0.0000	ワイブル分布	位置=55.35 尺度=140.56 形状=3.34	83.0860	0.2444
	正規分布	平均=111.58 標準偏差=31.76	499.4950	0.0000	正規分布	平均=181.51 標準偏差=41.62	149.3354	0.0000
V2	ベータ分布	α=5.58 β=4.51 尺度=224.51	401.6930	0.0000	ベータ分布	α=8.34 β=6.71 尺度=350.15	104.4002	0.0141
	正規分布	平均=124.07 標準偏差=33.52	724.3606	0.0000	ワイブル分布	位置=63.58 尺度=145.36 形状=3.30	175.3580	0.0000
	三角分布	最小=45.27 最尤値=118.45 最大=219.14	787.6080	0.0000	正規分布	平均=193.98 標準偏差=43.45	193.0856	0.0000
V3	ワイブル分布	位置=124.29 尺度=69.63 形状=2.46	156.3980	0.0000	ワイブル分布	位置=144.24 尺度=122.53 形状=3.25	90.7174	0.1043
	ガンマ分布	位置=66.04 尺度=6.06 形状=19.79	252.4146	0.0000	正規分布	平均=254.07 標準偏差=37.15	105.3482	0.0146
	対数正規分布	平均=186.04 標準偏差=26.96	264.2962	0.0000	ベータ分布	α=31.72 β=68.93 尺度=806.21	107.4180	0.0084

表２．５

	ディスカウントファクター 0.9				ディスカウントファクター 0.95			
		パラメータ他	χ二乗検定	P値		パラメータ他	χ二乗検定	P値
V1	ベータ分布	α=17.52 β=15.57 尺度=727.75	64.8054	0.7936	ベータ分布	α=28.88 β=25.10 尺度=1327.08	70.1300	0.6376
	正規分布	平均=385.22 標準偏差=62.21	75.8654	0.4828	正規分布	平均=710.11 標準偏差=89.27	80.7950	0.3317
	ワイブル分布	位置=144.09 尺度=264.62 形状=4.39	84.0972	0.2209	ワイブル分布	位置=338.28 尺度=406.18 形状=4.75	144.8482	0.0000
V2	ベータ分布	α=16.81 β=13.63 尺度=719.82	88.5686	0.1353	ベータ分布	α=28.02 β=23.04 尺度=1316.56	85.7246	0.1864
	正規分布	平均=397.50 標準偏差=63.83	99.9920	0.0340	正規分布	平均=722.48 標準偏差=90.81	88.6792	0.1515
	ワイブル分布	位置=120.15 尺度=302.16 形状=4.97	142.4782	0.0000	ガンマ分布	位置=-197.81 尺度=9.20 形状=100.00	159.0050	0.0000
V3	ベータ分布	α=31.59 β=35.49 尺度=967.91	111.2258	0.0042	ベータ分布	α=40.13 β=40.21 尺度=1560.59	63.7784	0.8190
	正規分布	平均=455.82 標準偏差=58.55	117.3404	0.0017	正規分布	平均=779.60 標準偏差=86.52	67.6652	0.7416
	ガンマ分布	位置=-133.60 尺度=5.89 形状=100.00	132.1134	0.0001	ガンマ分布	位置=-93.73 尺度=8.73 形状=100.00	114.6860	0.0022

３．確率最大化

ここまでの結果から、すなわち総報酬関数の分布が左右対称でもなく、また期待値基準で最適である政策による総報酬関数が都合のよい方に偏っているとは限らないことから、総報酬関数の分布をみることの有用性がわかる。ここでもう一つの例を見てみよう。この例では政策Ａと政策Ｂの違いは状態１における戦略の違いのみである。政策Ａに関するデータとその政策の下での割引期待総報酬は表３.１のとおりであり、政策Ｂに関するデータと割引期待総報酬は表３.２のとおりである。また、政策Ａ、政策Ｂそれぞれによる総報酬関数の分布の統計量とパーセンタイルは表３.３、表３.４のとおりである。

状態１についてのみ考える。期待値を用いる基準では政策Ｂが、政策Ａより良い。しかし政策Ａが次の期に70%の確率で90を達成する一方、政策Ｂでは90を達成するのに数期必要とする。政策Ａ、政策Ｂによる状態１の総報酬関数の分布の図はそれぞれ図３.５，図３.６のようになる。

これらをみると、政策Ａは値10から90までのどの値でも、その値を超える確率が70%で達成させることができることがわかる。さらに、これらの図とパーセンタイルから、59.04から90の間の値では政策Ａの方が基準値を超える達成確率が高く、91を超えて初めて政策Ｂが優位になることがわかる。期待値の違いもそれほど違いがなく、その期待値を達せいするまで政策Ｂは数期かかることから、政策Ａを選ぶことが考えられる。閾値を超える確率を最大にするような政策の選択が考えられる。

表３．１（政策Ａデータ）

推移確率				直接報酬	割引期待総報酬	
	状態1	状態2	状態3	(すべての状態で)	discount factor	0.8
状態1	0.0	0.7	0.3	10.0	状態1	66.0
状態2	0.0	0.0	1.0	100.0	状態2	100.0
状態3	0.0	0.0	1.0	0.0	状態3	0.0

表３．２（政策Ｂデータ）

推移確率				直接報酬	割引期待総報酬	
	状態1	状態2	状態3	(すべての状態で)	discount factor	0.8
状態1	0.9	0.0	0.1	20.0	状態1	71.4
状態2	0.0	0.0	1.0	100.0	状態2	100.0
状態3	0.0	0.0	1.0	0.0	状態3	0.0

表３．３（政策Ａ：統計量・パーセンタイル）

統計量:	値
試行回数	15000
平均値	66
中央値	90
最頻値	90
標準偏差	36.66
分散	1,344.09
歪度	−0.87
尖度	1.76
変動係数	0.56
範囲下限	10
範囲上限	90
範囲	80
平均標準誤差	0.3

パーセンタイル	値
0%	10
10%	10
20%	10
30%	10
40%	90
50%	90
60%	90
70%	90
80%	90
90%	90
100%	90

表３．４（政策Ｂ：統計量・パーセンタイル）

統計量:	値
試行回数	15000
平均値	71.31
中央値	79.03
最頻値	20
標準偏差	26.35
分散	694.15
歪度	−0.66
尖度	2.13
変動係数	0.37
範囲下限	20
範囲上限	99.99
範囲	79.99
平均標準誤差	0.22

パーセンタイル	値
0%	20
10%	20
20%	48.8
30%	59.04
40%	67.23
50%	79.03
60%	86.58
70%	93.13
80%	97.19
90%	99.26
100%	99.99

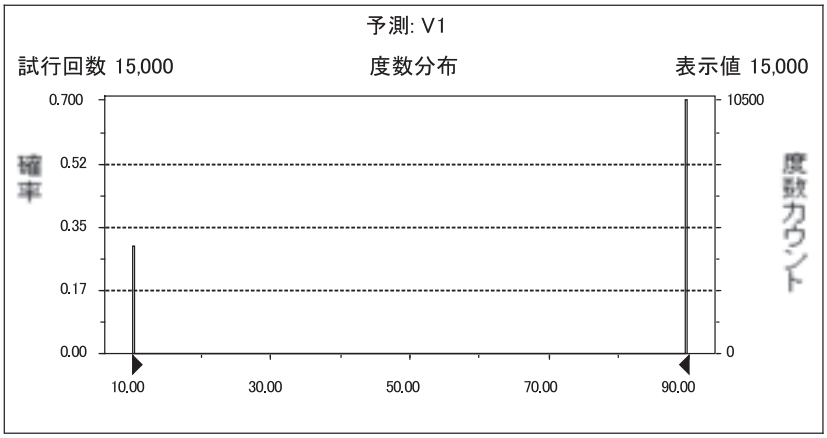

図３．５　政策Ａによる総報酬の分布

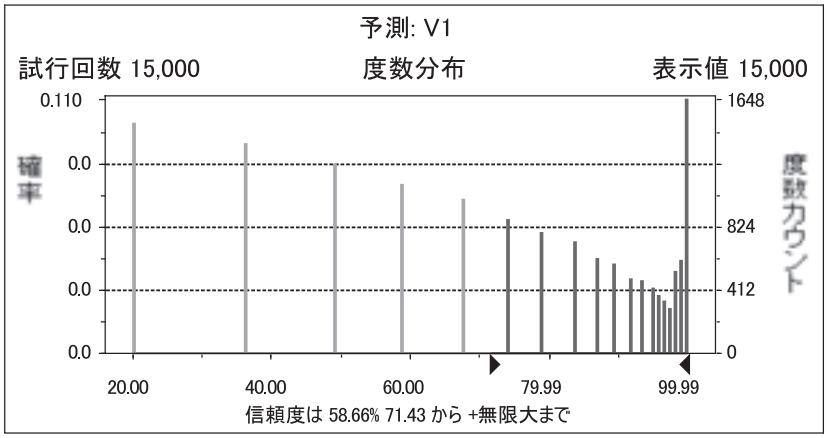

図３．６　政策Ｂによる総報酬の分布

４．終わりに

割引総報酬関数の分布を調べたが、クリスタルボールでは簡単にいろいろ変数の値をかえてそれによって変化する目的値の分布を図で見ることができる。推移確率行列も組み込みの分布を利用してスプレッドシート上に簡単に表現できた。本稿が、推移確率行列を用いるモデル分析の参考になれば幸いである。

参考文献

[1] Ｒ．Ａ．ハワード、ダイナミックプログラミングとマルコフ過程、培風館（1971）

Ronald A. Howard, *Dynamic Programming and Markov Processes,* (The Massachusetts Institute of Technology Press, U.S.A., 1960)

[2] 船木洋一、相補的なポジティヴDPとネガティヴDP、「人文社会論叢」(社会科学篇), 第13号,237-243(2005)

[3] 船木洋一、閾値確率基準マルコフ決定過程とポジティブDP、「人文社会論叢」(社会科学篇), 第11号,1-10(2004)