Ⅴ　参考・・・主成分分析による年齢階層別の差異（全世帯）

（１）分析の目的

平成 18 年における全世帯の年齢階層別 10 大費目に対する支出をもとに年齢階層別の差異について、多変量解析の１つである主成分分析の手法により分析する。

（２）分析対象データ

食料、住居、光熱･水道、家具･家事用品、被服及び履物、保健医療、交通･通信、教育、教養娯楽、その他の消費支出について、年齢階層別の各 10 個のデータによる。（表１）

（３）分析手法

標準化したデータに対して主成分分析を適用する相関係数行列による。

（注）データの単位や桁数が異なる場合にこの影響を排除するため、全てのデータを平均０、標準偏差１に変換することをデータの標準化という。具体的には、N 個のデータ「X_1、X_2、・・・、X_N」において、平均 $\overline{X}$、標準偏差 ν とした場合、変数 X のデータから、次式により平均０、標準偏差１の新しいデータ H_i を作ることである。

$$H_i = (X_i - \overline{X}) / \nu$$

（４）分析結果

①情報説明力

表２から、第 1 主成分は全 10 項目の情報量の 41.7%、第 2 主成分は 27.5%を説明している。第 2 主成分までで全体の 69.3%を説明したこととなる。

②軸の解釈

図１において、横軸が第 1 主成分を意味し、縦軸が第 2 主成分を意味する。表３から、第１主成分について 10 大費目の係数をみると、プラスの符号で比較的大きいのが、食料、光熱・水道、教育、教養娯楽、被服及び履物、マイナスの符号で比較的大きいのが住居（中年以上で持ち家の取得となるので逆に作用する。）である。衣食住と光熱・水道の基礎的消費支出に加えて教育が増加し、住居が持家となり減少するのは、子育ての進行時期と重なり合う。このことから、この軸は、子育てに沿った基礎的消費支出に関連していると考えられる。第 2 主成分について 10 大費目の係数をみると、プラスの符号で比較的大きいのが、保健医療、家事･家具用品、その他の消費支出、マイナスの符号が比較的大きいのが、交通･通信である。高齢になるにともない保健医療やその他消費支出（交際費）が大きくなり、また、活発な行動

力が減退してくる。このことから、この軸は高齢化に関連していると考えられる。

③年齢階層別の分類

図 1 において、30 歳未満、30～34 歳、35 歳～39 歳は第 3 象限にある。この年齢層は、子育ての初期であると考えられることから横軸が比較的小さく、また、行動が活発であると考えられることから縦軸が比較的小さい。

40～44 歳、45～49 歳、50～54 歳は第 4 象限にある。この年齢層は、子育ての盛期であると考えられることから横軸は比較的大きく、また、依然として行動の活発さを保持していると考えられることから縦軸が比較的小さい。

55～59 歳、60～64 歳は第 1 象限にある。この年齢階層は、子育ての終期と自身の豊かな生活への欲求が相まった状況にあると考えられることから横軸が比較的大きく、また、健康保持と交際に生活の中心を移行させたと考えられることから縦軸が比較的大きい。

65 歳以上は第 2 象限にある。この年齢階層は、退職（引退）後の年金等での 2 人暮らしの状況を反映したものであると考えられることから横軸が比較的小さく、健康保持と交際が生活の中心であると考えられることから縦軸が比較的大きい。

（注）主成分分析

１．主成分分析の意味

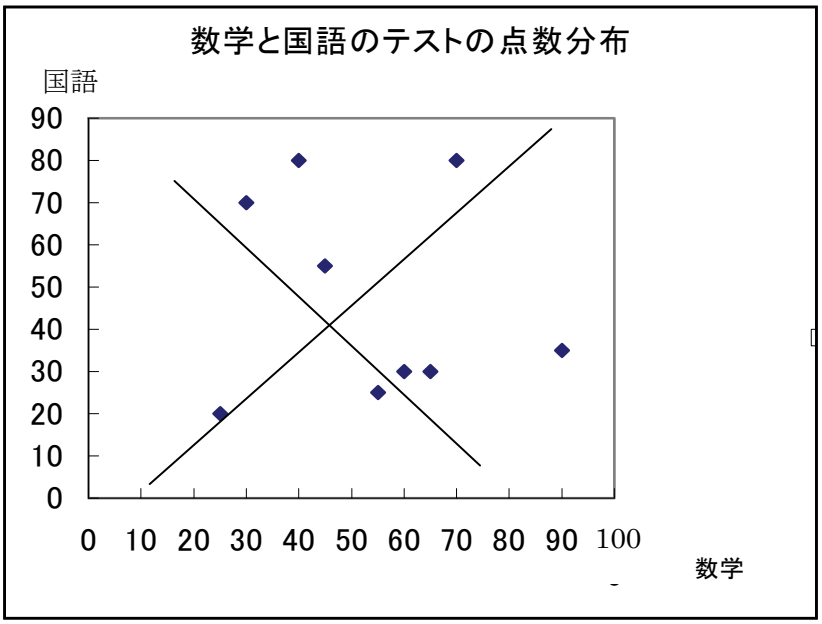
①ねらい

主成分分析は、ある対象について観測された多次元のデータを個々に分析するのではなく、そこに含まれる情報の損失を出来るだけ少なくして 2 次元や３次元にデータを縮約する手法である。主成分分析を活用すると、それらの項目全体が織りなす意味合いを解釈することができる。主成分分析は多変量解析の一つの手法である。

なお、個々の項目ごとに平均や分散を求めるなどデータを個々に分析する方法は、各項目のもつ意味を解析するものであり、一変量分析である。

②変数統合と主成分分析

例えば、学生の国語と数学のテストの点数について分析する。散布図上に右上がりの直線を引いてみる。この直線を軸と考えて目盛りを入れると、この軸は生徒の総合的な成績の良さを示していると考えられよう。さらに、この直線に垂直な直線を引く。これは数学が得意の数理能力に優れているか国語が得意の言語能力に優れているかという型を示すものと考えられよう。主成分分析はこのような軸を求めるための手法で、データを統合して新しい総合的な変数を作り出すことを目的としている。



２．主成分分析の手法

変数の数がｋ個（X_1，X_2，・・・，X_k）、観測対象の数がｎ個（回答者がｎ人）の多変量データがあるとする。

このデータをもとにｋ個より少ないｊ個の新しい変数Z_1、Z_2、・・・、Z_jを作り出すことを考える。

新しい変数Z_1、Z_2、・・・、Z_jは、もとの変数X_1、X_2、・・・、X_kを結合した変数で、次のような式で表せるようにしたい。

$Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1k}X_k$

$Z_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2k}X_k$

・・・

$Z_j = a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jk}X_k$

計算により求めたいのは、X_1、X_2、・・・、X_kの各係数a_{11}、a_{12}、・・・、a_{jk}である。

ここで、新しい変数は、次のような性質を持ったものにしたい。

①Z_1はX_1からX_kの情報が最大限集約されるようにする。（ｋ個の変数が持っている情報を１つの変数に集約しようとする場合には情報の損失が生じる。）

②Z_2はX_1、X_2、・・・、X_kの情報がZ_1の次にできるだけ多く集約されるようにする。しかも、Z_1とは独立になるようにする。

③Z_3はX_1、X_2、・・・、X_kの情報がZ_1とZ_2の次にできるだけ多く集約されるようにする。しかも、Z_1およびZ_2とは独立になるようにする。

④以下、Z_4からZ_jまで同様に考える。

このような性質を満足するようにa_{11}、a_{12}、・・・、a_{jk}を算出するのが主成分分析の計算である。

さて、①はZ_1の分散が最大になるようにすることと同じ意味をもつ。ところが、そのために、a_{11}、a_{12}、・・・、a_{1k}を限りなく大きくすればよく、それではZ_1が定まらない。そこで、

$$a_{11}^2 + a_{12}^2 + \cdots + a_{1k}^2 = 1$$

という条件をつける。

②はZ_1とは独立で、かつ、分散が最大になるようにすることと同じ意味をもつ。この場合も、

$$a_{21}^2 + a_{22}^2 + \cdots + a_{2k}^2 = 1$$

という条件をつける。

③と④も同様に考える。

このような条件のもとで、a_{11}、a_{12}、・・・、a_{jk}を求めることは、X_1、X_2、・・・、X_kの分散共分散行列の固有値と固有ベクトルを計算する（これは標準化しない場合であり、標準化する場合は相関係数行列の固有値、固有ベクトルを計算する）ことに帰

着し、a_{11}、a_{12}、・・・、a_{jk}は固有ベクトルにほかならない。

さて、新変数 Z_1、Z_2、・・・、Z_jの式が求まれば、その式に X_1、X_2、・・・、X_k の具体的な数値を代入することで、観測対象ごとに新変数の値を求めることができる。この数値のことを主成分スコアーと呼ぶ。

なお、多変量データは各変数が同じ単位で測定されている場合と変数の単位が不ぞろいの場合とがある。変数の単位が不ぞろいというのは、身長はｃｍの単位で測定され、体重はｋｇの単位で測定されている場合である。このような場合、変数ごとにデータを標準化してから、主成分分析を適応するほうがよい。なぜなら、主成分分析は測定単位のとり方に影響を受けるからである。変数ごとにデータを標準化することによって、変数間の単位の相異が解消できる。

３．主成分分析の種類

データを標準化せずに直接、原データに対して主成分分析を適用する方法を分散共分散行列から出発する主成分分析といい、標準化したデータに対して主成分分析を適用する方法を相関係数行列から出発する主成分分析という。どちらの行列から出発するかの判断基準は次のように考える。

・各変数の測定単位が異なるなどによるばらつきの違いを反映させたくない　⇒　相関係数行列

・各変数のばらつきの違いを反映させたい⇒　分散共分散行列

参考資料

内田　治「すぐわかる SPSS によるアンケートの多変量解析」　東京図書　2003 年

上田　尚一「統計用語辞典」　東洋経済新報社　1981 年

田中　豊、脇本　和昌「多変量統計解析法」　現代数学社　1983 年

表1　東京都の10大費目についての年齢階層別の支出（全世帯）　　単位：円

	食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
30歳未満	61,944	59,650	12,941	8,447	16,394	12,862	42,288	6,070	31,992	71,119
30～34歳	63,072	34,072	16,212	12,441	17,332	18,362	47,077	9,872	36,472	48,142
35～39歳	71,280	31,530	18,630	11,834	18,095	11,396	38,973	14,625	41,956	46,668
40～44歳	82,940	32,634	21,762	10,185	18,461	10,708	38,015	33,610	44,610	52,790
45～49歳	90,626	25,080	23,692	11,288	20,780	12,640	38,992	47,893	47,109	68,271
50～54歳	87,168	20,941	23,646	9,582	20,434	13,335	43,354	45,197	41,277	80,416
55～59歳	90,298	26,912	25,729	10,057	17,365	12,990	37,305	21,776	36,284	95,190
60～64歳	81,117	21,952	22,455	14,521	18,133	18,377	29,626	5,632	42,742	92,552
65歳以上	72,938	23,508	21,481	10,341	10,368	17,262	23,947	2,236	33,021	65,116

表2　主成分の情報吸収量

全情報量	90

	第1	第2	第3	第4	第5	第6	第7	第8
固有値	4.1737	2.7514	1.5312	0.8834	0.4696	0.1441	0.0427	0.0039
情報吸収量	37.5634	24.7625	13.7811	7.9504	4.2263	1.2971	0.3842	0.0349
比率	0.4174	0.2751	0.1531	0.0883	0.0470	0.0144	0.0043	0.0004
累積比率	0.4174	0.6925	0.8456	0.9340	0.9809	0.9953	0.9996	1.0000

表3　主成分の係数

	第1	第2	第3	第4	第5	第6	第7	第8
食料	0.4631	0.0873	0.2186	0.0267	0.0349	0.0251	0.4677	0.0961
住居	−0.3388	−0.3545	0.1220	0.1932	0.4464	−0.2551	0.6213	0.0115
光熱・水道	0.4191	0.2480	0.2003	−0.0736	−0.1469	0.3280	0.3619	−0.2645
家具・家事用品	0.0763	0.3203	−0.6460	0.1067	0.2626	0.2524	0.2042	0.5129
被服及び履物	0.3283	−0.3275	−0.2759	0.3750	0.1529	0.0207	−0.2749	−0.0923
保健医療	−0.1706	0.4353	−0.2898	0.3303	−0.4463	−0.5032	0.2184	−0.2441
交通・通信	0.0007	−0.4981	−0.2073	0.3915	−0.4483	0.3735	0.1943	−0.1416
教育	0.4050	−0.2817	0.0679	−0.0684	−0.3184	−0.5081	−0.0030	0.5476
教養娯楽	0.3979	−0.0962	−0.3654	−0.2139	0.3336	−0.3275	−0.0046	−0.5186
その他の消費支出	0.1734	0.2631	0.3724	0.7012	0.2698	−0.0668	−0.2480	0.0259