요인분석(Factor Analysis)

1. 개념

많은 변수들의 상호관련성을 이용하여 각 변수들의 잠재하고 있는 공통적인 요인(factor)을 찾아내어 이를 바탕으로 전체자료의 특성,특징을 설명하는 통계적 기법이며 이를 요인분석이라 한다.

2. 목적

여러 변수들이 가지고 있는 데이터를 이용하여 보다 적은 개수의 변수로 축소하여 전체 자료를 설명하는 것이 목적이다.

3. 전제조건

1) 모든 변수는 등간척도 이상의 연속형 변수로 측정이 되어야 하며,
2) 각 변수는 서로 독립이며 정규분포, 등분산을 이루고 있어야 하고,
3) 표본의 수는 50이상 내지 변수의 5배수 이상을 권고하고 있다.

4. 주요용어

1) 요인(factor) : 상관관계가 높은 변수끼리 묶어 유사한 속성을 가진 하나의 계층으로 묶어 놓은 것
2) 고유값(eigen value) : 추출된 요인의 고유값을 의미하며 일반적으로 고유값의 크기가 1 이상[1]인 요인을 기준으로 설명한다. 요인적재값의 제곱의 합으로서 크면 클수록 추출된 요인이 변수들의 분산을 잘 설명하고 있다는 것을 의미한다.
3) 공통성(communality) : 추출된 요인이 변수들의 속성을 얼마나 잘 반영하고 있는가에 대한 설명력을 의미한다. 1을 기준으로 소수점이하 셋째자리로 출력이 된다. A변수의 요인에 대한 공통성이 0.815라면 A변수의 변동중 81.5%가 요인에 의해 설명되고 있다는 의미이다.
4) 요인적재값(factor loading) : 변수들과 요인간의 상관계수를 의미하며 0.3이상이면 일반적인 유의성을, 0.5이상이면 높은 유의성이 있다고 해석한다.
5) 요인행렬(factor matrix) : 요인에 대한 변수들의 요인적재값을 모아놓은 것
6) 요인회전(factor rotation) : 추출된 요인행렬을 해석이 용이하게 회전하는 것을 의미하여 quartimax, varimax, equamax, oblimin, promaxe 등의 5가지 옵션이 있는데 사회과학분야에서처럼 변수들이 완전한 독립을 이루지 못하는 경우의 요인회전 방법으로 oblimin회전 방법을 주로 이용한다.
7) 요인점수(factor score) : 각 case(표본)의 요인점수를 산출한다.

---

1) 고유값의 기준을 1이상인 요인으로 설명하는 경우에 1개의 요인만 설명이 되는 경우가 많아 0.7~0.8이상의 고유값을 기준으로 요인을 설명하기도 한다.

5. 요인분석과정

다음은 어느 **시의 행정서비스만족도에 대한 조사결과이다. 이 데이터를 이용해 요인분석을 하기로 한다.

1) 분석-데이터축소-요인분석

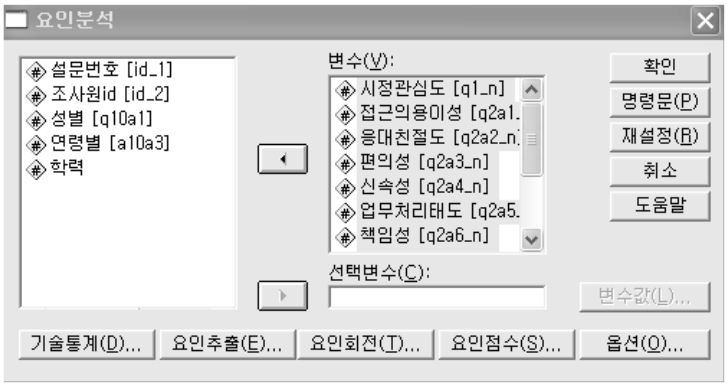

그림 1-1.요인분석대화상자

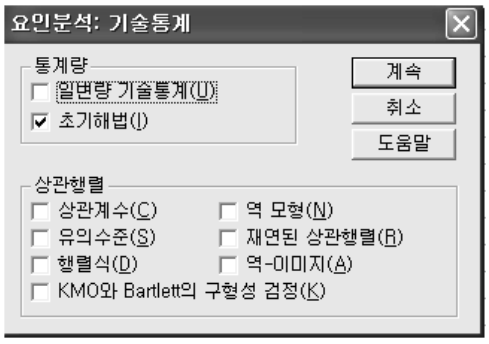

그림 1-2 기술통계 옵션

디폴트값은 '초기해법'에만 체크가 되어있다. 통계량의 탭에서는 '일변량기술통계', 상관행렬 탭에서는 '상관계수'에 체크한다.

KMO는 표본의 수와 변수의 수가 적당한가에 대한 검정통계량으로 KMO>0.8이면 good으로 판정하는 것이 일반적이다.

Bartlett 검정은 요인분석에서 사용하는 상관계수의 행렬의 대각행렬이라는 귀무가설에 대한 검정이다. 대각행렬은 곧 변수(독립변수)간의 상관관계가 없음을 의미하는 것으로 Bartlett 검정통계량이 0.05이하이면 요인분석이 가능한 데이터구조라고 판정한다.

2) 요인추출

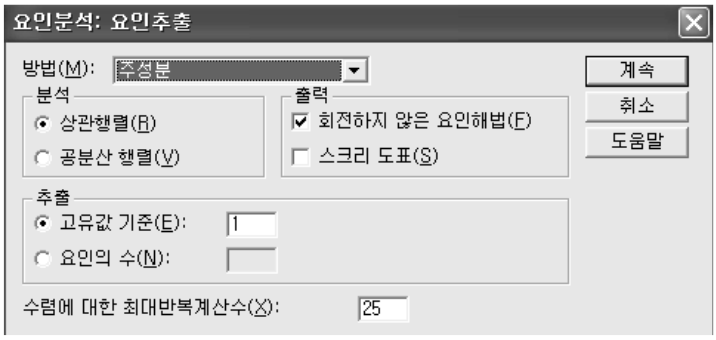

그림 1-3 요인추출방법 대화상자

요인추출에서는 추출방법은 디폴트의 '주성분'을 선택하며,
분석의 탭에서는 '상관행렬'과 '공분산행렬'의 2가지가 있는데 측정된 변수들의 단위를 고려[2]하여 선택해야 한다.
출력의 탭에서는 스크리도표를 체크해줌으로서 추출해야할 요인의 개수를 정하는 것이 용이해진다.
추출의 탭에서는 디폴트로 고유값이 1이 되어 있다.

3) 요인회전

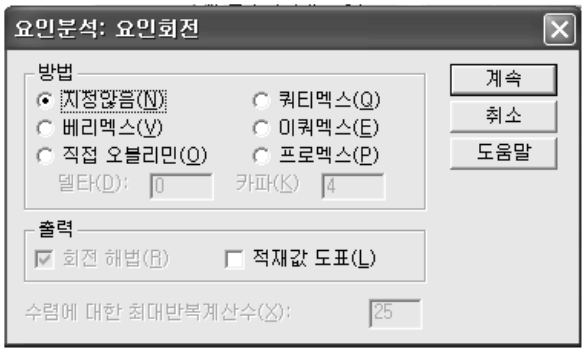

그림 1-4 요인회전 대화상자

방법의 탭의 디폴트는 지정되지 않았다. 5가지의 요인회전방법중 사회과학분야에서 주로 사용하는 '직접 오블리민[3]'에 체크를 해주고 '적재값도표'에도 체크해준다.

---

2)변수들의 측정단위가 다른 경우=상관행렬,
변수들의 측정단위가 동일한 경우=공분산행렬
3)변수간에 상호 독립이라는 기본 가정을 충실할 수 없는 사회과학분야에서 주로 사용하는 방식

4) 요인점수

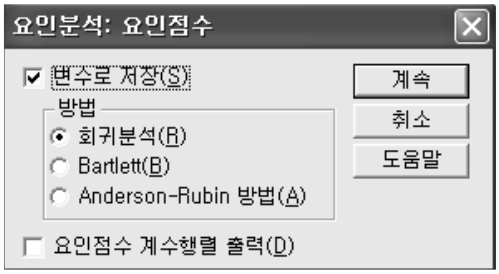

그림 1-5 요인점수 대화상자

요인점수의 탭에서는 '변수로 저장'과 방법은 '회귀분석'를 선택한다.

5) 요인분석 '옵션'

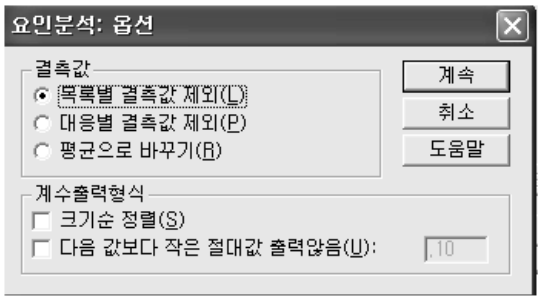

그림 1-6 요인분석 '옵션'

요인분석의 옵션에서는 결측값의 지정방식과 계수출력에 대한 세부적인 사항을 선택할 수 있다.

6. 요인분석결과의 해석

출력결과는 다음과 같이 제시된다.

기술통계량

	평균	표준편차	분석수
시정관심도	3.30	.880	765
접근의용이성	3.49	.902	765
응대친절도	3.36	1.041	765
편의성	3.31	.923	765
신속성	3.32	.944	765
업무처리태도	3.17	.923	765
책임성	3.21	.960	765
쾌적성	3.68	.868	765
신뢰성	3.22	.918	765
체감만족도	3.00	.710	765

표 1-7 기술통계량

상관행렬

		시정관심도	접근의용이성	응대친절도	편의성	신속성	업무처리태도	책임성	쾌적성	신뢰성	체감만족도
상관계수	시정관심도	1.000	.170	.126	.138	.117	.113	.085	.136	.115	.166
	접근의용이성	.170	1.000	.658	.619	.642	.554	.543	.507	.521	.361
	응대친절도	.126	.658	1.000	.598	.631	.613	.569	.510	.599	.411
	편의성	.138	.619	.598	1.000	.855	.615	.560	.476	.612	.336
	신속성	.117	.642	.631	.855	1.000	.624	.591	.514	.604	.346
	업무처리태도	.113	.554	.613	.615	.624	1.000	.697	.498	.692	.392
	책임성	.085	.543	.569	.560	.591	.697	1.000	.524	.681	.373
	쾌적성	.136	.507	.510	.476	.514	.498	.524	1.000	.576	.338
	신뢰성	.115	.521	.599	.612	.604	.692	.681	.576	1.000	.467
	체감만족도	.166	.361	.411	.336	.346	.392	.373	.338	.467	1.000

표 1-8 상관행렬 출력결과

상관행렬의 출력결과를 보고 각 변수들간의 상관관계를 보며 추출한 요인의 형태를 가늠하게 된다. '접근의 용이성', '응대친절도'간(0.658), '책임성', '업무처리태도'간(0.697),'편의성', '신속성'간에 높은 (+)적 상관관계(0.855)가 있으며, '이외에도 높은 상관관계를 보이는 변수가 보이고 있다. 요인분석은 여러 변수들중 전체자료를 설명할 수 있는 요인으로 묶어주는 것이 분석의 목적이므로 대략 2~3개 정도의 요인이 적당할 것이라는 추측을 하게 된다.

공통성

	초기	추출
시정관심도	1.000	.922
접근의용이성	1.000	.609
응대친절도	1.000	.648
편의성	1.000	.686
신속성	1.000	.721
업무처리태도	1.000	.682
책임성	1.000	.650
쾌적성	1.000	.495
신뢰성	1.000	.686
체감만족도	1.000	.378

추출 방법: 주성분 분석.

표 1-9 공통성 출력결과

공통성 출력결과는 '추출된 요인이 변수들의 속성을 얼마나 잘 반영하고 있는가에 대한 기여도'를 의미한다. '시정관심도'변수의 요인에 대한 공통성이 0.922을 나타내고 있다. '시정관심도'변수의 변동중 92.2%가 요인에 의해 설명되고 있다는 의미이다.

설명된 총분산

	초기 고유값			추출 제곱합 적재값			회전 제곱합 적재값[a]		
성분	전체	% 분산	% 누적	전체	% 분산	% 누적	전체	% 분산	% 누적
1	5.473	54.735	54.735	5.473	54.735	54.735	5.466		
2	1.003	10.029	64.764	1.003	10.029	64.764	1.295		
3	.794	7.939	72.703						
4	.628	6.276	78.979						
5	.561	5.607	84.586						
6	.483	4.826	89.412						
7	.348	3.479	92.891						
8	.288	2.882	95.773						
9	.283	2.831	98.604						
10	.140	1.396	100.000						

추출 방법: 주성분 분석.

a. 성분이 상관된 경우 전체 분산을 구할 때 제곱합 적재값이 추가될 수 없습니다.

표 1-10 주성분분석으로 추출된 요인들의 고유값과 분산

'%분산'이란 추출된 요인에 의해 설명될 수 있는 분산비율을 요인별로 나열을 하고 있다. 요인이 추가될수록 분산비가 높아지며 분산의 제곱근이 설명력을 의미한다. 표에서는 고유값1 이상인 요인을 2개로 추출하고 있으며 이들의 기여도는 64.76%라는 결과를 출력하고 있다.

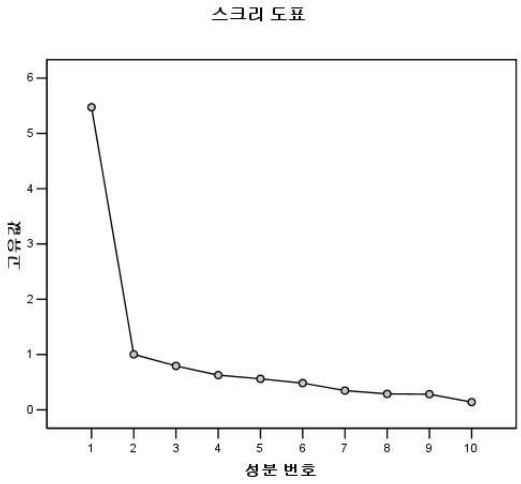

그림 1-11 스크리도표

스크리도표는 고유값의 크기를 좌에서 우로 나열하고 있는데 경사가 거의 완만하게 이루어지기 시작하는 구간의 성분의 수는 2~3을 나타내고 있다. 다시 말해 2~3의 요인이 전체 변수들을 설명하는데 적당할 것이라는 추가 정보를 제공하는 것이다.
사람의 팔 꺾이는 부분을 연상하면 이해가 쉬울 것이다.

성분행렬[a]

	성분	
	1	2
시정관심도	.199	.939
접근의용이성	.780	.033
응대친절도	.805	-.023
편의성	.824	-.079
신속성	.843	-.103
업무처리태도	.821	-.090
책임성	.797	-.121
쾌적성	.703	.023
신뢰성	.827	-.050
체감만족도	.549	.277

요인추출 방법: 주성분 분석.
a. 추출된 2 성분

패턴 행렬[a]

	성분	
	1	2
시정관심도	-.049	.969
접근의용이성	.763	.067
응대친절도	.803	.011
편의성	.837	-.046
신속성	.861	-.069
업무처리태도	.836	-.057
책임성	.821	-.090
쾌적성	.690	.054
신뢰성	.831	-.016
체감만족도	.471	.306

요인추출 방법: 주성분 분석.
회전 방법: Kaiser 정규화가 있는 오블리민.
a. 3 반복계산에서 요인회전이 수렴되었습니다.

표 1-12 성분/패턴행렬

요인적재량을 파악할 수 있는 표를 출력하고 있다. 성분행렬표가 요인회전전의 요인적재량을, 패턴행렬은 요인회전후의 요인적재량을 나타내고 있는데 일반적으로 0.3이상이면 추출된 요인이 통계적으로 의미가 있으며, 0.5이상이면 매우 유의한 것으로 해석한다.
요인분석결과 시정관심도, 체감만족도를 제외한 8개변수가 요인1에 의해 설명되는 요인적재량이 매우 높다는 결론을 내리게 된다.

패턴행렬표를 보며 10개의 변수들이 가지고 있는 특성을 요약해서 설명을 할 수 있어야 하는데, 시정관심도와 체감만족도를 제외한 8개의 변수의 요인적재량이 모두 높아 요약하기가 다소 곤란한 느낌이 드는 요인분석결과라고 봐야겠다. 굳이 요약해서 요인의 이름을 명명한다면 요인1은 행정서비스의 인식요인, 요인2는 시정관심요인으로 분류할 수 있겠다.
요인을 명명[4]하는 것은 추출된 요인을 연구결과를 받아들이는 대중에게 좀 더 간결하고 함축적으로 전달하기 위한 일련의 과정인 것이다. 요인2은 시정관심도 변수를 제외한 모든 요인적재량이 0.5미만으로 명명하기가 보다 명확하게 보여지고 있다.
실제 요인분석결과를 해석하는데 있어 요인의 이름을 명명하는 것은 대단히 중요하고 함축적인 의미로 전체자료를 설명할 수 있는 것이라고 봐야한다. 명명의 방법에 대한 구체적이고 체계적인 방법은 연구자 본인이 스스로 정해야 하는 것이고, 그에 대한 최소한의 변수 구분 요건은 **요인적재량(factor loading) >0.5** 이 된다.

다음은 요인점수(factor score)를 이용하여 두 요인을 산점도를 이용하여 시각화 해보기로 한다.

spss메뉴-그래프-선도표-다중도표를 선택한 후 '도표에 표시할 데이터'의 탭은 '개별변수의 요약값'으로 하여 추출된 2요인을 연령별로 시각화를 하기로 하자.

그림 1-13 선도표 대화상자

4) 명명방법에 대한 일반적인 정의는 없으며, 연구자의 연구의 목적, 취지등에 비추어 명명하는 것이 보편적이다.

그림 1-14 다중선도표 정의 대화상자

그래프

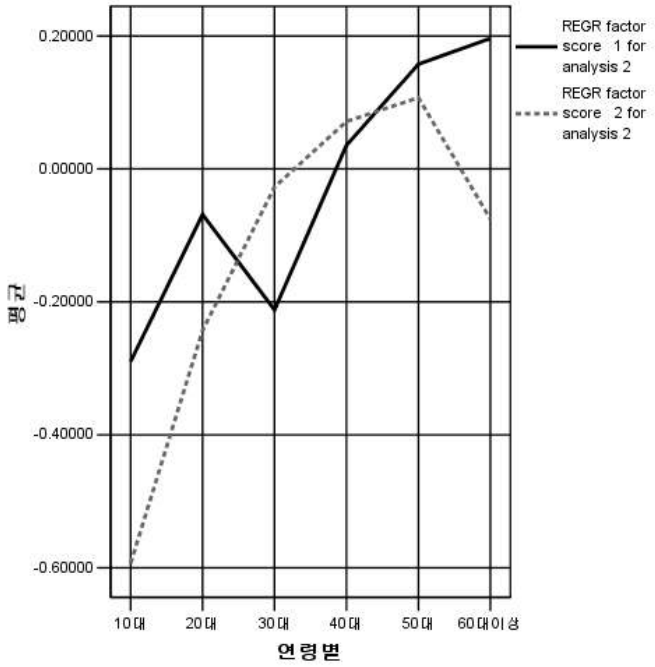

그림 1-15 선도표 출력결과

그림1-15에서 보듯 10대,60대에서 두 요인간의 점수차가 크게 나타나고 있음을 알 수 있으며 이에 대한 통계적인 평균차이[5]를 대응표본t검정을 통해 제시할 수도 있다.

5) 두가지 요인은 서로 대응(동일한 객체)하고 있기 때문이다.

spss메뉴-데이터-파일분할에서 '집단들비교'를 선택하고 분할 집단변수란에 연령변수를 투입한다.

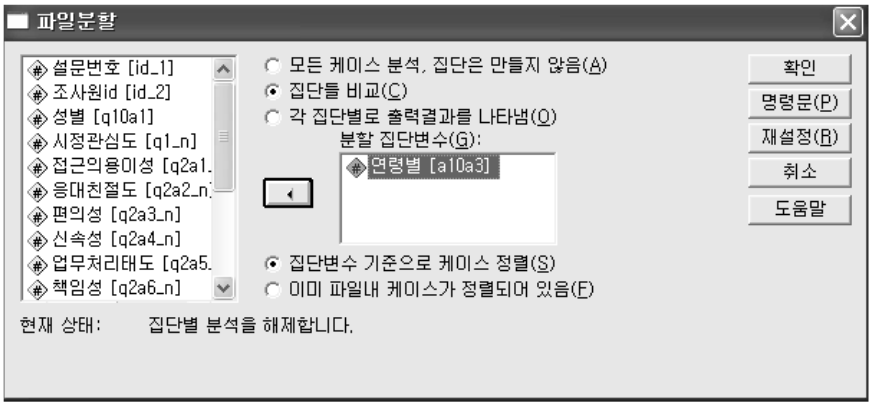

그림 1-16 파일분할 대화상자

대응표본검정

연령별			대응차					t	자유도	유의확률 (양쪽)
			평균	표준편차	평균의 표준오차	차이의 95% 신뢰구간 하한	상한			
10대	대응 1	REGR factor score 1 for analysis 2 - REGR factor score 2 for analysis 2	.30364155	.88834822	.22208706	-.16972580	.77700891	1.367	15	.192
20대	대응 1	REGR factor score 1 for analysis 2 - REGR factor score 2 for analysis 2	.17432432	1.4051231	.17039620	-.16578793	.51443657	1.023	67	.310
30대	대응 1	REGR factor score 1 for analysis 2 - REGR factor score 2 for analysis 2	-.18501054	1.2183243	.08815485	-.35889847	-.01112260	-2.099	190	.037
40대	대응 1	REGR factor score 1 for analysis 2 - REGR factor score 2 for analysis 2	-.03582545	1.2667170	.08109325	-.19556085	.12390994	-.442	243	.659
50대	대응 1	REGR factor score 1 for analysis 2 - REGR factor score 2 for analysis 2	.05034589	1.2126564	.09219656	-.13163649	.23232827	.546	172	.586
60대이상	대응 1	REGR factor score 1 for analysis 2 - REGR factor score 2 for analysis 2	.27167378	1.2630436	.14885112	-.02512692	.56847449	1.825	71	.072

그림 1-17 대응표본검정 출력결과

추출된 2요인을 다중선도표로 출력하면 요인간의 평균차이를 가늠하게 되지만 통계적인 평균차이는 대응표본검정결과 30대와 60대이상의 연령대가 유의수준 0.1에서 유의하게 나타났다고 결론을 내린다.

7. 논문, 보고서 제시요령

요인명	변수명	Mean[6]	SD[7]	Factor[8] loading	eigen[9] value	Variance[10] (%)
요인1 (**********) α=0.***				0.82	4.325	53.42
				0.54		
				0.84		
				0.67		
				0.80		
				0.68		
				0.58		
				0.61		
				0.69		
				0.92		
요인2 (**********) α=0.***				0.69	1.235	
				0.69		
				0.92		

* 요인적재량이 +0.5 이상인 변수의 통계량을 입력한다.

6) 평균
7) 표준편차
8) 요인적재값
9) 고유값
10) 분산