요인분석과 상관분석

요인분석

1) 요인분석 개요
2) 공통요인으로 변수 정제
3) 잘못 분류된 요인 제거로 변수 정제
4) 요인분석 결과 제시

요인분석 개요

- **요인분석(Factor Analysis)**
  - 다수의 변수들을 대상으로 변수들 간의 관계 분석(타당성 분석)
  - 공통 차원으로 축약하는 통계기법(변수 축소)
  - 탐색적 요인분석 : 요인분석을 할 때 사전에 어떤 변수들끼리 묶어야 한다는 전제를 두지 않고 분석하는 방법
  - 확인적 요인분석 : 요인분석을 할 때 사전에 묶여질 것으로 기대되는 항목끼리 묶여지는지를 분석하는 방법

요인분석 개요

- **요인과 변수 관계**

요인 구분	변수명(Name)	변수설명(하위 요인)
제품 친밀도	q1	브랜드
	q2	친근감
	q3	익숙함
	q4	편안함
제품 적절성	q5	가격의 적절성
	q6	당도의 적절성
	q7	성분의 적절성
제품 만족도	q8	음료의 목 넘김
	q9	음료의 맛
	q10	음료의 향
	q11	음료의 가격

요인분석 개요

【요인분석의 전제조건】

- 하위요인으로 구성되는 데이터 셋이 준비되어 있어야 한다.
- 분석에 사용되는 변수는 등간척도나 비율척도이여야 하며, 표본의 크기는 최소 50개 이상이 바람직하다.【중심극한정리】
- 요인분석은 상관관계가 높은 변수들끼리 그룹화하는 것이므로 변수들 간의 상관관계가 매우 낮다면(보통 ±3 이하) 그 자료는 요인 분석에 적합하지 않다.

요인분석 개요

【요인분석의 목적】

- 자료 요약 : 변인을 몇 개의 공통된 변인으로 묶음
- 변인 구조 파악 : 변인들의 상호관계 파악(독립성 등)
- 불필요한 변인 제거 : 중요도가 떨어진 변수 제거
- 측정도구 타당성 검증 : 변인들이 동일한 요인으로 묶이는지 여부를 확인

공통요인으로 변수 정제

【데이터 셋 준비】

```
s1 <- c(1, 2, 1, 2, 3, 4, 2, 3, 4, 5)
s2 <- c(1, 3, 1, 2, 3, 4, 2, 4, 3, 4)
s3 <- c(2, 3, 2, 3, 2, 3, 5, 3, 4, 2)
s4 <- c(2, 4, 2, 3, 2, 3, 5, 3, 4, 1)
s5 <- c(4, 5, 4, 5, 2, 1, 5, 2, 4, 3)
s6 <- c(4, 3, 4, 4, 2, 1, 5, 2, 4, 2)
name <-1:10
```

s1 : 자연과학, s2 : 물리화학

s3 : 인문사회, s4 : 신문방송

s5 : 응용수학, s6 : 추론통계

name : 각 과목 문항 이름

공통요인으로 변수 정제

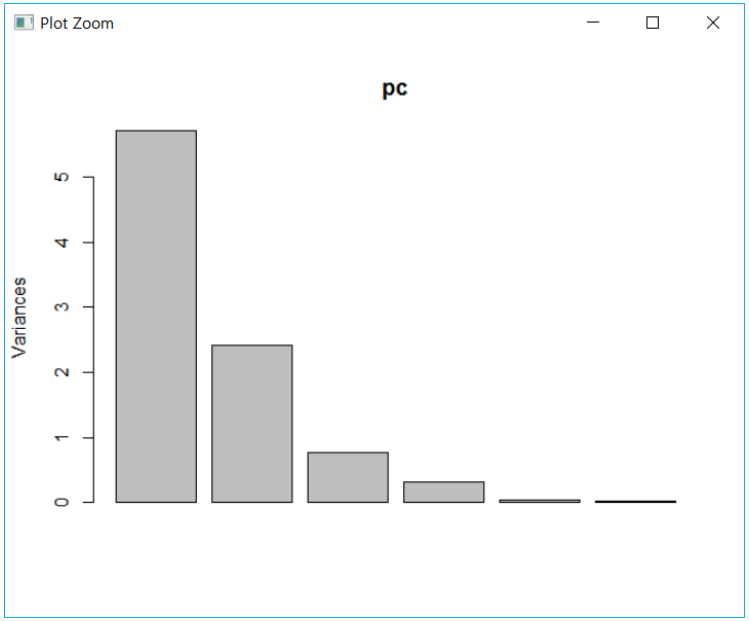
【주성분 분석】

- 변동량(분산)에 영향을 주는 주요 성분을 분석하는 방법.
- 요인 분석에서 사용될 요인의 개수를 결정하는데 주로 이용.

【주성분분석 요인 수 분석】

- 요인분석에서 공통 요인으로 묶일 요인 수를 알아본다.

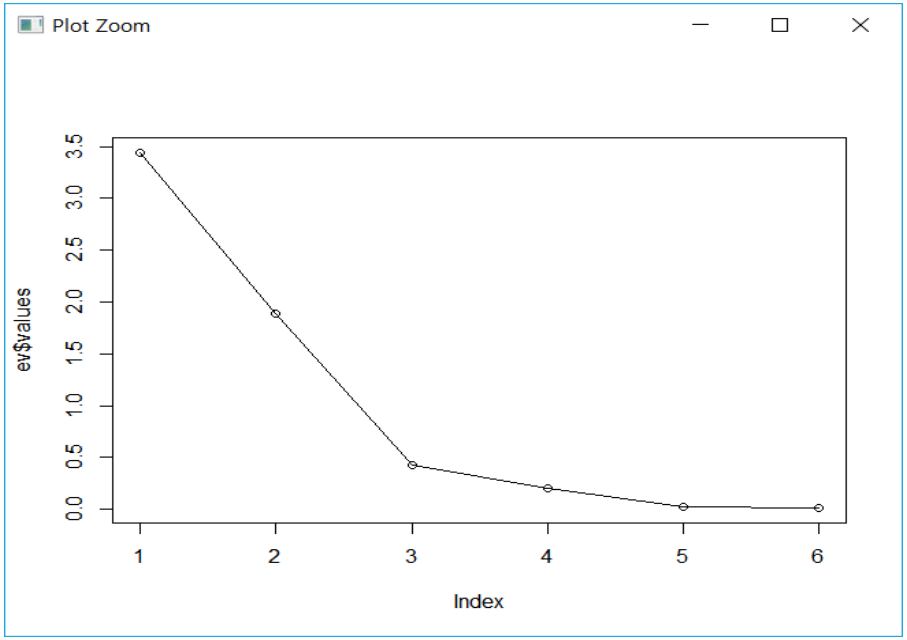
```
pc <- prcomp(subject) # 주성분분석 수행 함수
summary(pc) # 요약통계량
plot(pc)
```



공통요인으로 변수 정제

【주성분분석 요인 수 분석】

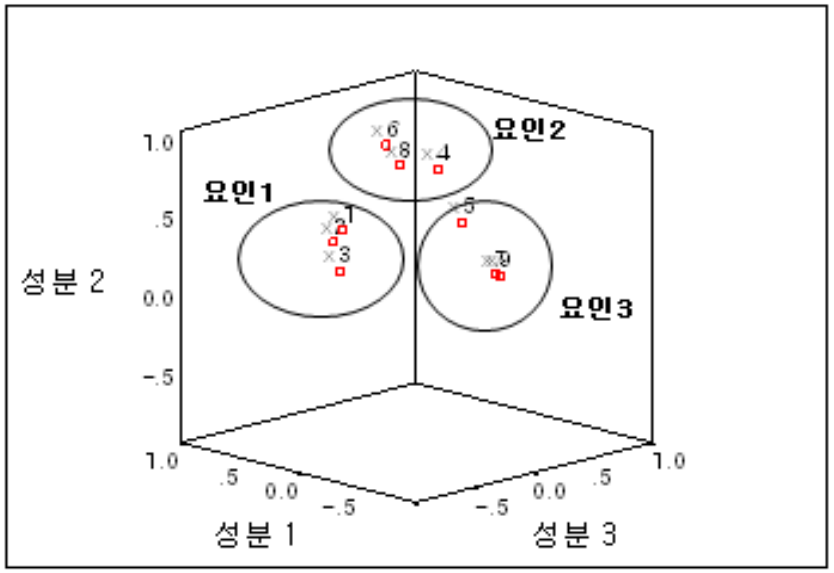
```
# 초기 고유값 계산
en <- eigen(cor(subject))
#$values : 고유값 보기
en$values
# $vectors : 고유벡터 보기
en$vectors
# 고유값을 이용한 시각화
plot(ev$values, type="o")
```



공통요인으로 변수 정제

【요인분석 : 베리멕스 직각회전법 적용】

```
result <- factanal(subject, factors = 3, # 요인 개수 지정
                   rotation = "varimax", # 회전방법 지정("varimax", "promax", "none")
                   scores="regression") # 요인점수 계산 방법
```



[베리멕스 직각회전법 예]

공통요인으로 변수 정제

【요인점수】

- 각 관측치(표준화된 값)와 요인 간의 관계를 통해서 구해진 점수

subject

	s1	s2	s3	s4	s5	s6
1	1	1	2	2	4	4
2	2	3	3	4	5	3
3	1	1	2	2	4	4
4	2	2	3	3	5	4
5	3	3	2	2	2	2
6	4	4	3	3	1	1
7	2	2	5	5	5	5
8	3	4	3	3	2	2
9	4	3	4	4	4	4
10	5	4	2	1	3	2

+

요인적재량

Loadings:

	Factor1	Factor2	Factor3
S1	-0.379	-0.005	0.923
s2	-0.710	0.140	0.649
s3	0.236	0.931	0.166
s4	0.120	0.983	-0.118
s5	0.771	0.297	-0.278
s6	0.900	0.301	-0.307

=

요인점수

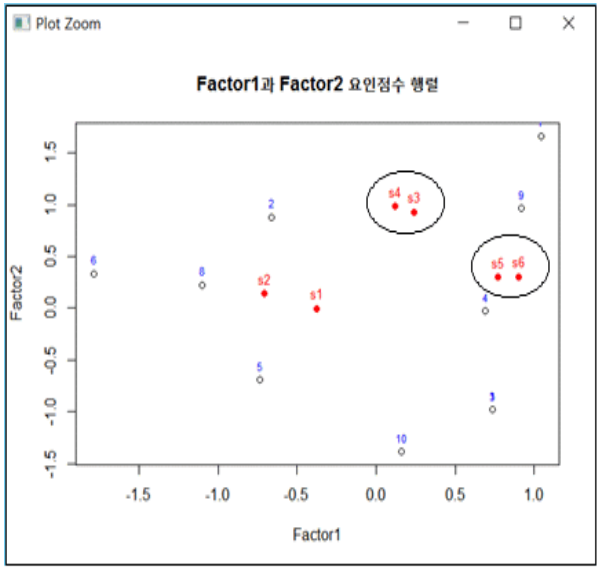
	Factor1	Factor2	Factor3
[1,]	0.7357870	-0.98034177	-1.07981805
[2,]	-0.6640013	0.87937769	-0.83543481
[3,]	0.7357870	-0.98034177	-1.07981805
[4,]	0.6917075	-0.02812698	-0.27885523
[5,]	-0.7387206	-0.69135360	-0.07138837
[6,]	-1.7858690	0.33608991	0.30957945
[7,]	1.0449596	1.66369477	-0.11745856
[8,]	-1.0999660	0.22263533	-0.17382007
[9,]	0.9197524	0.96404108	1.40734566
[10,]	0.1605633	-1.38567464	1.91966803

공통요인으로 변수 정제

【요인점수를 이용한 요인적재량 시각화】

단계 1 : Factor1과 Factor2 요인적재량 시각화

```
plot(result$scores[, c(1:2)],
        main="Factor1과 Factor2 요인점수 행렬")
# 산점도에 레이블 표시(문항 이름 : name)
text(result$scores[,1], result$scores[,2],
labels = name, cex = 0.7, pos = 3, col = "blue")
# 요인적재량 추가
points(result$loadings[,c(1:2)], pch=19, col = "red")
 # 요인적재량의 레이블 표시
text(result$loadings[,1], result$loadings[,2],
                        labels = rownames(result$loadings),
                        cex = 0.8, pos = 3, col = "red")
```

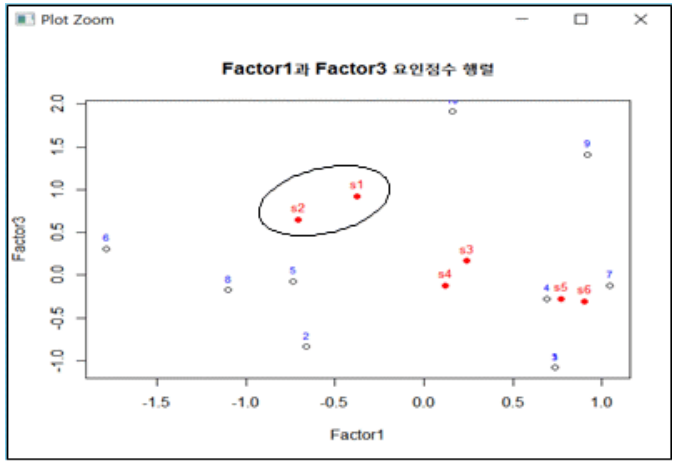


공통요인으로 변수 정제

【요인점수를 이용한 요인적재량 시각화】

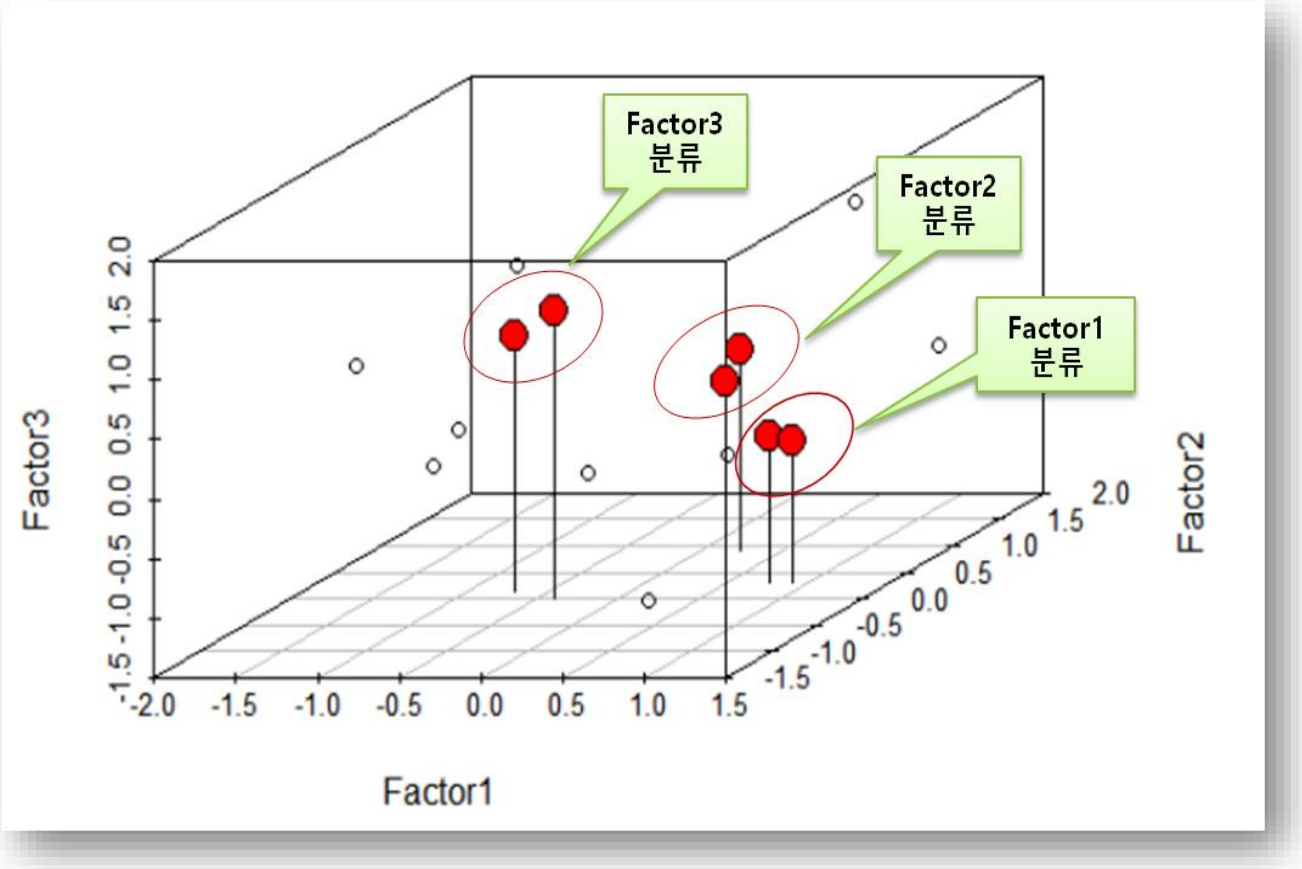
단계 2 : Factor1과 Factor3 요인적재량 시각화

```
plot(result$scores[,c(1,3)], main="Factor1과 Factor3 요인점수 행렬")
# 산점도에 레이블 표시(문항 이름 : name)
text(result$scores[,1], result$scores[,3],
                        labels = name, cex = 0.7, pos = 3,
                        col = "blue")
# 요인적재량 추가
points(result$loadings[,c(1,3)], pch=19, col = "red")
 # 요인적재량의 레이블 표시
text(result$loadings[,1], result$loadings[,3],
                        labels = rownames(result$loadings),
                        cex = 0.8, pos = 3, col = "red")
```



공통요인으로 변수 정제

【요인점수를 이용한 요인적재량 3차원 시각화】



잘못된 요인 정제

【 3개 요인으로 구성된 파일정보(drinking_water.sav)】

요인 구분	변수명(Name)	변수설명(하위 요인)	변수값(Values)
제품 친밀도	q1	브랜드	**5점 척도** ① 매우불만 ② 불만 ③ 보통 ④ 만족 ⑤ 매우만족 (무응답 없음)
	q2	친근감	
	q3	익숙함	
	q4	편안함	
제품 적절성	q5	가격의 적절성	
	q6	당도의 적절성	
	q7	성분의 적절성	
제품 만족도	q8	음료의 목 넘김	
	q9	음료의 맛	
	q10	음료의 향	
	q11	음료의 가격	

요인분석 결과 제시 방법

【논문/보고서 작성방법】

요인 (Factor)	변수명 (Variable Name)	요인 적재량 (Factor loading)	고유값 (Eigenvalue)	분산 설명력 (Variance Explained)
제품 친밀도	q1	.762	2.133	19.4%
	q2	.813		
	q3	.762		
제품 적절성	q5	.557	2.394	21.8%
	q6	.693		
	q7	.703		
제품 만족도	q8	.695	2.772	25.2%
	q9	.873		
	q10	.852		
	q11	.719		

상관분석

1) 상관분석 개요
2) 피어슨 상관계수
3) 상관분석 실습
4) 상관분석 결과 제시

상관분석 개요

상관관계 분석(Correlation Analysis)

- 변수 간 관련성 분석 방법
- 하나의 변수가 다른 변수와 관련성 분석
- 예) 광고비와 매출액 사이의 관련성, 광고량과 브랜드 인지도 등 분석

【상관관계분석 중요사항】

- 회귀분석 전 변수 간 관련성 분석(가설 검정 전 수행)
- 상관계수 ➔ **피어슨(Pearson) R계수** 이용 관련성 유무
  - ✓ 상관관계분석 척도:
  - ✓ 피어슨 상관계수(Pearson correlation coefficient : r)

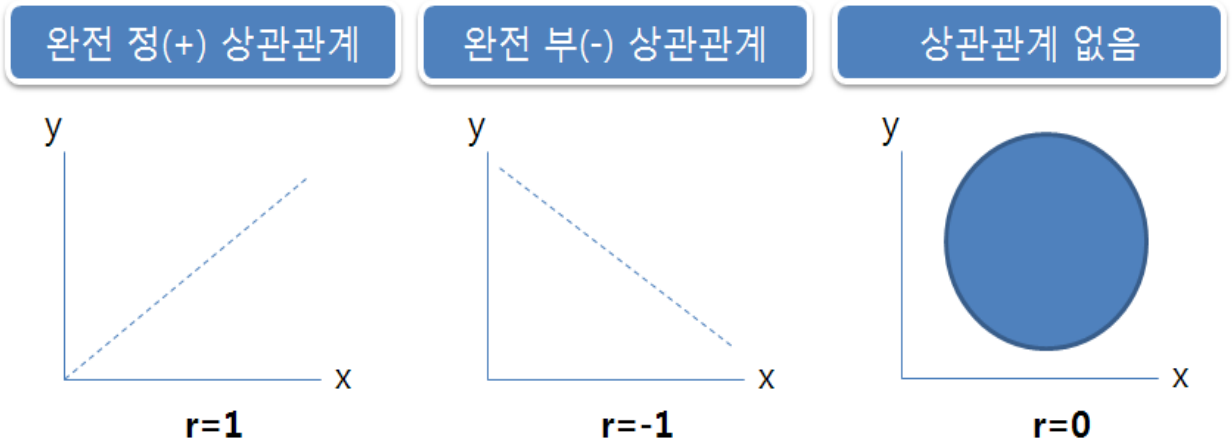
피어슨 상관계수 R

【피어슨 상관계수 R】

피어슨 상관계수 R	상관관계 정도
±0.9이상	매우 높은 상관관계
±0.9 ~ ±0.7	높은 상관관계
±0.7 ~ ±0.4	다소 높은 상관관계
±0.4 ~ ±0.2	낮은 상관관계
±0.2 미만	상관관계 없음
※ 상관계수 r은 -1에서 +1까지의 값을 가진다. 또한 가장 높은 완전 상관관계의 상관계수는 1이고, 두 변수간에 전혀 상관관계가 없으면 상관계수는 0이다.	

피어슨 상관계수 R

- **상관계수 r과 상관관계 정도**



상관분석 실습

```
# 데이터셋 가져오기
result <- read.csv("C:/Rwork/data/drinking_water.csv", header=T)
head(result)

# 상관계수 보기
cor(result$친밀도, result$적절성)
cor(result$친밀도, result$만족도)

# 전체 변수 간 상관계수 보기
cor(result, method="pearson") # 피어슨 상관계수 - default

cor(result, method="spearman") # spearman 상관계수(서열척도)
```

상관분석 결과 제시

【논문에서 상관관계 분석 결과 제시 방법】

– 일반적으로 상관관계 분석 결과를 논문에서 제시할 경우 해당 기술통계량(평균과 표준편차)과 피어슨 상관계수 함께 제시

분석 단위	평균 (Mean)	표준편차 (Std. Deviation)	분석 단위 간 상관관계 (Inter-Analysis Correlations)		
			1	2	3
1. 친밀도	2.928	0.9703446	1		
2. 적절성	3.133	0.8596574	.499	1	
3. 만족도	3.095	0.8287436	.467	.767	1