요인분석 (Factor Analysis)

http://contents2.kocw.or.kr/KOCW/document/2018/wonkwang/chunghoil0415/14.pdf
http://contents.kocw.or.kr/KOCW/document/2015/chungbuk/najonghwa1/7.pdf

ch14. 요인분석, 상관분석

요인분석: 변수들의 상관성을 바탕으로 변수를 정제하여 상관관계 분석이나 회귀분석에서 설명변수(독립변수)로 사용된다.

1. 요인분석

요인분석(Factor Analysis): 다수의 변수를 대상으로 변수 간의 관계를 분석하여 공통 차원으로 축약하는 통계기법

요인분석
1) 탐색적 요인분석: 요인 분석을 할 때 사전에 어떤 변수들끼리 묶어야 한다는 전제를 두지 않고 분석하는 방법
2) 확인적 요인 분석: 사전에 묶일 것으로 기대되는 항목끼리 묶였는지를 조사하는 방법

타당성: 측정 도구가 측정하고자 하는 것을 정확히 측정할 수 있는 정도
논문 작성을 위한 통계분석 방법에서 인구통계학적 분석(빈도분석, 교차분석 등)을 시행한 이후 통계량 검정 이전에 구성 타당성(Construct validity)검증을 위해서 요인분석(Factor Analysis) 시행

요인분석을 위한 전제조건
- 하위요인으로 구성되는 데이터 셋이 준비되어 있어야 한다.
- 분석에 사용되는 변수는 등간척도나 비율척도여야 하며, 표본의 크기는 최소 50개 이상이 바람직하다.
- 요인 분석은 상관관계가 높은 변수들끼리 그룹화하는 것이므로 변수 간의 상관관계가 매우 낮다면(보통 ±3이하), 그 자료는 요인 분석에 적합하지 않다.

요인분석을 수행하는 목적
- 자료의 요약: 변인을 몇 개의 공통된 변인으로 묶음
- 변인 구조 파악: 변인들의 상호관계 파악(독립성 등)
- 불필요한 변인 제거: 중요도가 떨어진 변수 제거

■ 측정 도구의 타당성 검증: 변인들이 동일한 요인으로 묶이는 지를 확인

요인 분석 결과에 대한 활용방안

1. 타당성 검정: 측정도구가 정확히 측정했는지를 알아보기 위하여 측정변수들이 동일한 요인으로 묶이는지를 검정
2. 변수 축소: 변수들의 상관관계가 높은 것끼리 묶어서 변수를 정제
3. 변수 제거: 변수의 중요도를 나타내는 요인적재량이 0.4미만이면 설명력이 부족한 요인으로 판단하여 제거
4. 활용: 요인 분석에서 얻어지는 결과를 이용하여 상관분석이나 회귀분석의 설명변수로 활용

1.1 공통요인으로 변수 정제

특정항목으로 묶이는데 사용되는 요인 수 결정은 주성분 분석 방법과 상관계수 행렬을 이용한 초기 고유값을 이용

실습 (변수와 데이터프레임 생성)

1단계: 과목 변수 생성

```
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
```

2단계: 과목 데이터프레임 생성

```
subject <- data.frame(s1, s2, s3, s4, s5, s6)
str(subject)
```

실습 (변수의 주성분 분석)

주성분 분석: 변동량(분산)에 영향을 주는 주요 성분을 분석하는 방법

1단계: 주성분 분석으로 요인 수 알아보기

```
pc <- prcomp(subject)
summary(pc)
plot(pc)
```

```
prcomp(subject)
```

*주성분 분석에서 결정된 주성분의 수를 반드시 요인 분석에서 요인의 수로 사용되지는 않는다.

고유값: 어떤 행렬로부터 유도되는 실수값

일반적으로 변화량의 합(총분산)을 기준으로 요인의 수를 결정하는데 이용된다.

2단계: 고유값으로 요인수 분석

```
en <- eigen(cor(subject))
names(en)

en$values
en$vectors

plot(en$values, type = "o")
```

고유값이 급격하게 감소하다가 완만하게 감소할 때 급격하게 감소하는 고유값 index 수로 주성분 변수 개수 결정

실습 (변수 간의 상관관계 분석과 요인분석)

1단계: 상관관계분석 – 변수 간의 상관성으로 공통요인 추출

```
cor(subject)
```

요인분석에서 요인회전법은 요인 해석이 어려운 경우 요인축을 회전시켜서 요인 해석을 용이하게 하는 방법을 의미.
대표적인 요인회전법으로는 배리맥스 회전법

요인분석에 이용되는 R함수: factanal()함수

형식: factanal(dataset, factors=요인수, scores=c("none", "regression", "Bartlett"), rotation="요인회전법", ...)

2단계: 요인분석 – 요인회전법 적용(Varimax회전법)

2-1단계: 주성분 분석의 가정에 의해서 2개 요인으로 분석

```
result <- factanal(subject, factors = 2, rotation = "varimax")
result
```

* 요인 분석 결과에서 만약 p-value값이 0.05미만이면 요인수가 부족하다는 의미로 요인수를

늘려서 다시 분석을 수행해야 한다.

varimax 요인회전법 설명
직각회전
cf. 사각회전방식: promax

2-2단계: 고유값으로 가정한 3개 요인으로 분석

```
result <- factanal(subject,
                   factor = 3,
                   rotation = "varimax",
                   scores = "regression")
result
```

Result 해설:
Uniquenesses항목: 유효성을 판단하여 제시한 값으로 통상 0.5이하이면 유효한 것으로 본다.
Loadings 항목은 요인 적재값(Loading)을 보여주는 항목으로 각 변수와 해당 요인간의 상관관계 계수를 나타낸다.
요인적재값(요인부하량)이 통상 +0.4이상이면 유의하다고 볼 수 있다.
만약 +0.4미만이면 설명력이 부족한 요인(중요도가 낮은 변수)으로 판단할 수 있다.
요인적재값(Loading)이 높게 나타났다는 의미는 해당 변수들이 해당 요인으로 잘 설명된다는 의미
SS loadings 항목: 각 요인 적재값의 제곱의 합을 제시한 값. 각 요인의 설명력을 보여준다.
Proportion Var 항목: 설명된 요인의 분산 비율로 각 요인이 차지하는 설명력의 비율
Cumulative Var 항목: 누적 분산 비율. 요인의 분산 비율을 누적하여 제시한 값.
현재 정보손실은 1-0.94 = 0.06으로 적정한 상태. 만약 정보손실이 너무 크면 요인 분석의 의미가 없어진다.

3단계: 다양한 방법으로 요인적재량 보기

```
attributes(result)

result$loadings

print(result, digits = 2, cutoff = 0.5)

print(result$loadings, cutoff = 0)
```

실습 (요인점수를 이용한 요인적재량 시각화)

요인점수(요인 분석에서 요인의 추정된 값)를 얻기 위해서는 scores속성(scores="regression": 회귀분석으로 요인점수 계산) 설정

1단계: Factor1과 Factor2 요인적재량 시각화

```
plot(result$scores[ , c(1:2)],
     main = "Factor1과 Factor2 요인점수 행렬")

text(result$scores[ , 1], result$scores[ , 2],
     labels = name, cex = 0.7, pos = 3, col = "blue")
```

요인점수행렬의 산점도

2단계: 요인적재량 추가

```
points(result$loadings[ , c(1:2)], pch = 19, col = "red")

text(result$loadings[ , 1], result$loadings[ , 2],
     labels = rownames(result$loadings),
     cex = 0.8, pos = 3, col = "red")
```

3단계: Factor1과 Factor3 요인 적재량 시각화

```
plot(result$scores[ , c(1, 3)],
     main = "Factor1과 Factor3 요인점수 행렬")
text(result$scores[ , 1], result$scores[ , 3],
     labels = name, cex = 0.7, pos = 3, col = "blue")
```

4단계: 요인적재량 추가

```
points(result$loadings[ , c(1, 3)], pch = 19, col = "red")
text(result$loadings[ , 1], result$loadings[ , 3],
     labels = rownames(result$loadings),
     cex = 0.8, pos= 3, col = "red")
```

실습 (3차원 산점도로 요인적재량 시각화)

1단계: 3차원 산점도 패키지 로딩

```
library(scatterplot3d)
```

2단계: 요인점수별 분류 및 3차원 프레임 생성

```
Factor1 <- result$scores[ , 1]
Factor2 <- result$scores[ , 2]
Factor3 <- result$scores[ , 3]

d3 <- scatterplot3d(Factor1, Factor2, Factor3, type = 'p')
```

Scatterplot3d()함수: 3차원 산점도
형식: scatterplot3d(x축, y축, z축, type="p")

3단계: 요인적재량 표시

```
loadings1 <- result$loadings[ , 1]
loadings2 <- result$loadings[ , 2]
loadings3 <- result$loadings[ , 3]
d3$points3d(loadings1, loadings2, loadings3,
            bg = 'red', pch = 21, cex = 2, type = 'h')
```

d3$points3d(): scatterplot3d()함수 내에서 점 찍기

실습 (요인별 변수 묶기)

1단계: 요인별 과목 변수 이용 데이터프레임 생성

```
app <- data.frame(subject$s5, subject$s6)
soc <- data.frame(subject$s3, subject$s4)
nat <- data.frame(subject$s1, subject$s2)
```

2단계: 요인별 산술평균 계산

```
app_science <- round((app$subject.s5 + app$subject.s6) / ncol(app), 2)
soc_science <- round((soc$subject.s3 + soc$subject.s4) / ncol(soc), 2)
nat_science <- round((nat$subject.s1 + nat$subject.s2) / ncol(net), 2)
```

산술평균을 계산하여 요인별로 3개의 파생변수 생성

3단계: 상관관계분석

```
subject_factor_df <- data.frame(app_science, soc_science, nat_science)
cor(subject_factor_df)
```

요인분석을 통해서 만들어진 파생변수는 상관분석이나 회귀분석에서 독립변수로 사용할 수 있다.

1.2 잘못 분류된 요인 제거로 변수 정제

특정 변수가 묶여질 것으로 예상되는 요인으로 묶이지 않는 경우 해당 변수를 제거하여 변수를 정제하는 방법

실습 (요인분석에 사용될 데이터 셋 가져오기)

[표 14.1] 3개 요인으로 구성된 파일 정보(drinking_water.sav)

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

1단계: spss데이터 셋 가져오기

```
install.packages("memisc")
library(memisc)
setwd("C:/Rwork/ ")
data.spss <- as.data.set(spss.system.file('drinking_water.sav'))
data.spss[1:11]
```

memisc 패키지: spss데이터를 가져오기 위한 패키지
as.data.set(spas.system.file('spss데이터 파일')함수: spss데이터를 가져오기

2단계: 데이터프레임으로 변경

```
drinking_water <- data.spss[1:11]
drinking_water_df <- as.data.frame(data.spss[1:11])
str(drinking_water_df)
```

실습 (요인 수를 3개로 지정하여 요인 분석 수행)

```
result2 <- factanal(drinking_water_df, factor = 3, rotation = "varimax")
result2
```

Uniquenesses항목에서 모든 변수가 0.5이하의 값을 갖기 때문에 모두 유효
데이터 셋에서 의도한 요인과 요인으로 묶여진 결과가 서로 다르게 나옴
q4 변수의 타당성 의심

p-value < 0.05 → 요인변수 선택에 문제가 있음
여기서 p-value는 chi_square검정으로 기대치와 관찰치에 차이가 있음을 알려주는 확률값
요인 수를 3개로 전제하기 대문에 p-value값은 무시하고 실습 진행

실습 (요인별 변수 묶기)

1단계: q4를 제외하고 데이터프레임 생성

```
dw_df <- drinking_water_df[-4]
str(dw_df)
dim(dw_df)
```

2단계: 요인에 속하는 입력 변수별 데이터프레임 구성

```
s <- data.frame(dw_df$Q8, dw_df$Q9, dw_df$Q10, dw_df$Q11)
c <- data.frame(dw_df$Q1, dw_df$Q2, dw_df$Q3)
p <- data.frame(dw_df$Q5, dw_df$Q6, dw_df$Q7)
```

3단계: 요인별 산술평균 계산

```
satisfaction <- round(
  (s$dw_df.Q8 + s$dw_df.Q9 + s$dw_df.Q10 + s$dw_df.Q11) / ncol(s), 2)
closeness <- round(
  (c$dw_df.Q1 + c$dw_df.Q2 + c$dw_df.Q3) / ncol(s), 2)
pertinence <- round(
  (p$dw_df.Q5 + p$dw_df.Q6 + p$dw_df.Q7) / ncol(s), 2)
```

요인 제품만족도(satisifaction)를 구성하는 4개의 변수의 산술평균 계산
요인 제품친밀도(closeness)와 요인 제품만족도(pertinence)는 3개 변수에 대한 산술평균을 계산하여 파생변수를 생성

4단계: 상관관계 분석

```
drinking_water_factor_df <- data.frame(satisfaction, closeness, pertinence)
colnames(drinking_water_factor_df) <- c("제품만족도", "제품친밀도", "제품적절성")
cor(drinking_water_factor_df)

length(satisfaction); length(closeness); length(pertinence)
```

1.2 요인 분석 결과 제시

요인 분석 결과를 논문이나 보고서에 제시하는 경우, 해당 요인과 변수, 요인적재량, 요인의 설명력을 나타내는 고유값을 함께 제시

ch14 연습문제 1, 2번 풀기

실습.
http://contents.kocw.or.kr/KOCW/document/2015/chungbuk/najonghwa1/8.pdf