

## Day33; 20221021

	날짜
	유형
	태그

GitHub - u8yes/R

You can't perform that action at this time. You signed in with another tab or window. You signed out in another tab or window. Reload to refresh your session. Reload to refresh your session.

u8yes/R



<https://github.com/u8yes/R>

As 1

Contributor

0

Issues

1

Star

0

Forks



[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/a56e35c1-7764-4e3a-94d0-bc2ca8db8b47/chap13\\_Ttest\\_Anova.r](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/a56e35c1-7764-4e3a-94d0-bc2ca8db8b47/chap13_Ttest_Anova.r)

[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/316a9869-65ed-4d56-bf71-0c63300d3f57/chap14\\_Factor\\_Correlation\\_Analysis.r](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/316a9869-65ed-4d56-bf71-0c63300d3f57/chap14_Factor_Correlation_Analysis.r)

[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/99097dec-15c7-45b8-8ea5-1ccb90985f8b/14.%EC%9A%94%EC%9D%B8%EB%B6%84%EC%84%9D%EA%B3%BC\\_%EC%83%81%EA%B4%80%EB%B6%84%EC%94%9D.pdf](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/99097dec-15c7-45b8-8ea5-1ccb90985f8b/14.%EC%9A%94%EC%9D%B8%EB%B6%84%EC%84%9D%EA%B3%BC_%EC%83%81%EA%B4%80%EB%B6%84%EC%94%9D.pdf)

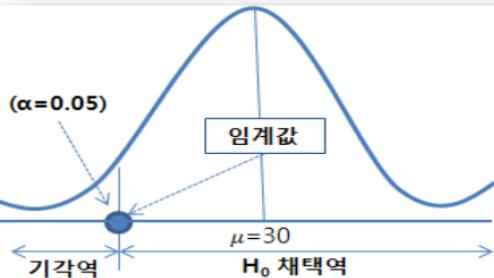
- 단측검정(1-sided test) : 방향(우열) 있는 단측가설 검정

$H_0$  : 1일 생산되는 불량품의 개수는 평균 30개 이다. ( $\mu=30$ )

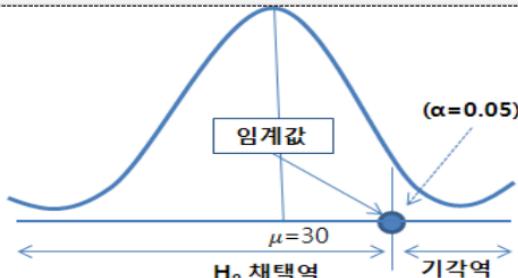
$H_1$  : 1일 생산되는 불량품의 개수는 평균 30개 이하이다. ( $\mu < 30$ ) ▶ 원쪽 단측검정

1일 생산되는 불량품의 개수는 평균 30개 이상이다. ( $\mu > 30$ ) ▶ 오른쪽 단측검정

연구가설이 < 또는 > 두 가지 가설 포함



A. 좌측검정



B. 우측검정

왼쪽 단측검정

오른쪽 단측검정

귀무가설일 경우에는 단측 검정은 아무 소용 없다. 연구 가설에 의해서 단측가설을 검정할 수 있는 것이다.

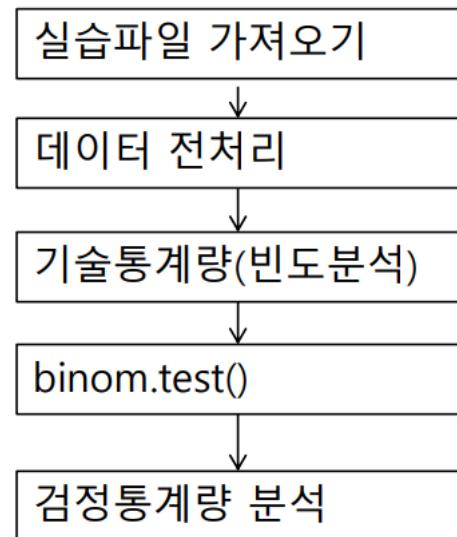
```
# 방향성을 갖는 단측 가설 검정
binom.test(14, 150, p = 0.2, alternative = "greater", conf.level = 0.95) # p-value = 0.9999
binom.test(14, 150, p = 0.2, alternative = "less", conf.level = 0.95) # p-value = 0.0003179 # 불만률은 낮아졌다라고 유의수준을 보고 판단할 수 있다
# 150명 중 14명이 불만을 나타냄
```

비율

# 단일집단 비율 검정

---

## ● 분석절차

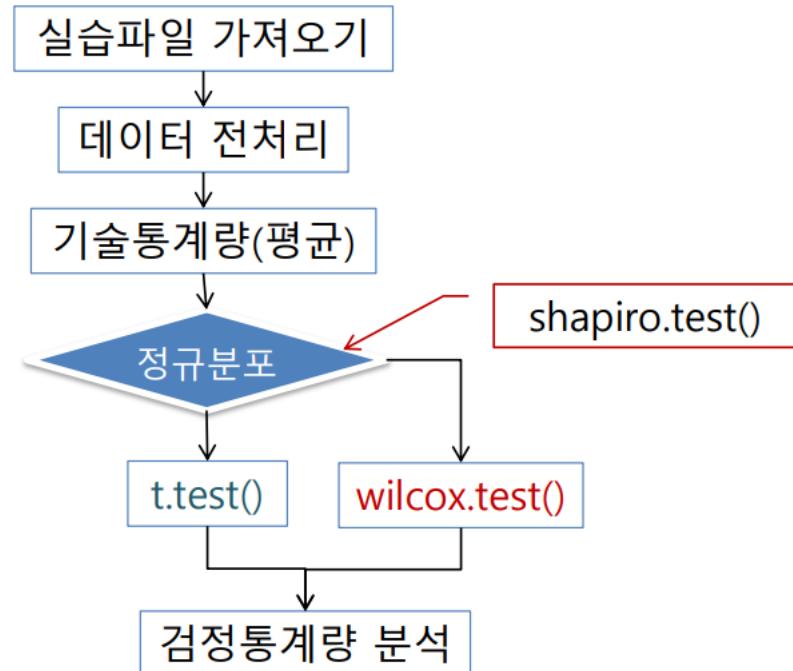


---

평균

# 단일집단 평균 검정

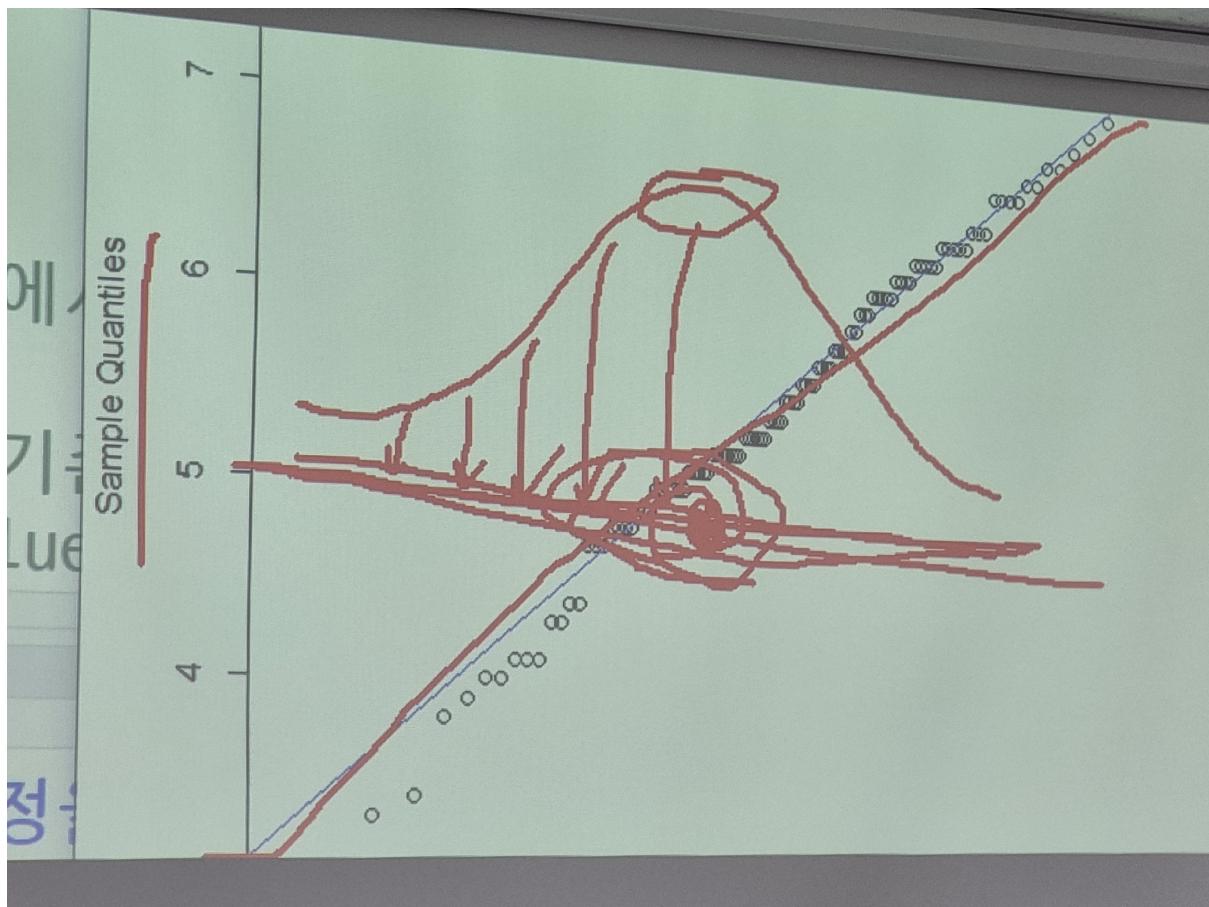
## ● 분석절차



```
# 단계4. 정규분포 검정
# 귀무가설(H0) : x의 데이터 분포는 정규분포이다.

shapiro.test(x1)
# 정규분포 검정 함수(p-value = 0.7242), 표본이 정규 분포로부터 추출된 것인지 테스트하기 위한 함수. 이때 귀무가설은 주어진 데이터가 정규 분포로부터의 표본이라는 것이다.
# p-value > α(일파) : 정규분포로 본다.
# 가설 설정이 아니기 때문에, p-value값으로 판단을 할 때, 귀무가설을 기준으로 보기 때문에 0.05보다 커야 정규분포로 본다.
```

```
# stats 패키지에서 정규성 검정을 위해서 제공되는 시각화 함수.
qqnorm(x1)
qqline(x1, lty=1, col='blue') # 선으로 정규성 여부를 파악하고자 함.
```



선을 대각선으로 기준으로 정규분포를 보면 된다.

동일한 표본(동일한 집단)은 대응 두 집단

## 대응 두 집단 평균 검정 (대응표본 T검정)

```
#####
# 추론통계학 분석 – 2-3. 대응 두 집단 평균 검정(대응표본 T검정)
#####
# 방법 : 동일한 표본을 대상으로 측정된 두 변수의 평균 차이를 검정하는
#         분석.
```

```

423 # (5) 사후검정
424 TukeyHSD(result) # 분산분석의 결과로 사후검정
425 plot(TukeyHSD(result))
426
423:11 (Untitled) ▾
Console Terminal × Background Jobs ×
R 4.1.3 · D:/heaven_dev/workspaces/R/data/ ↵
95% family-wise confidence level

Fit: aov(formula = score ~ method2, data = data2)

$method2
      diff      lwr      upr      p adj
방법2-방법1 2.612903 1.9424342 3.2833723 0.0000000
방법3-방법1 1.422903 0.7705979 2.0752085 0.0000040
방법3-방법2 -1.190000 -1.8656509 -0.5143491 0.0001911

```

차이 하한가 상한가 p-value관련(0.05보다 작은 값을 가지면 3집단 간 분산의 차이값이 의미가 있다)

---

### 요인분석

대표적인 비지도 학습방법의 알고리즘.

데이터 셋이 주어졌을 때 유사성을 검토해서 분류를 해줌.

## 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

탐색적 요인분석 - 아무런 사전지식을 부여하지 않는다.

요인 분석은 유사한 feature가 있다면 다 사용하는 것이 아닌 그 값을 대표값으로 표현할 수 있는 방법 강구.

---

시계열 - 다중공선성 등을 살펴보자.

---

알고리즘에는 연속형을 넣는 경우는 덜하다, 범주형으로 바꿔서 넣는다.

---

범주형과 이산형을 구분하기도 한다.

이산 - discrete.

범주 - 연속형 변수(예: 나이(청년, 중년, 장년층 등)를 역코딩해서 만든 것.

연속 - continuous.

---

```
#####
# Chapter14-1. 요인분석(Factor Analysis) # 요인 - domain, 값(빈도수를 가지고 있는 값)
#####

# 요인분석의 목적
# 1. 자료의 요약 : 변인(변수)을 몇 개의 공통된 변인으로 묶음
# 2. 변인 구조 파악 : 변인들의 상호관계 파악(독립성 등)
# 3. 불필요한 변인(변수) 제거 : 중요도가 떨어진 변수 제거
# 4. 측정도구 타당성 검증 : 변인(변수)들이 동일한 요인으로 묶이는지 여부를 확인

# 전제조건 : 등간척도 or 비율척도, 정규분포, 관찰치 상호독립적/분산 동일

# 요인분석 결과에 대한 활용 방안
# 1. 서로 밀접하게 관련된 변수들을 합치거나 중복된 변수를 제거하여 변수를 축소한다.
# 2. 변수들 간의 연관성 또는 공통점 탐색
# 3. 요인점수 계산으로 상관분석, 회귀분석의 설명변수로 이용
```

## 공통요인으로 변수 정제

---

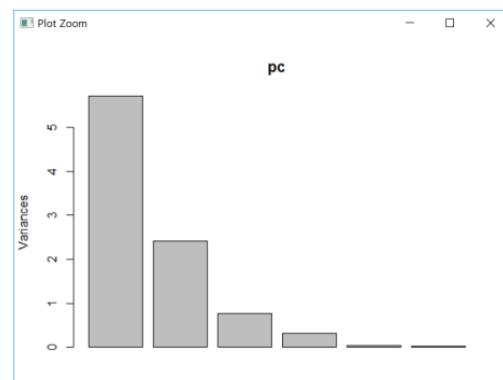
### 【주성분 분석】

- 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.

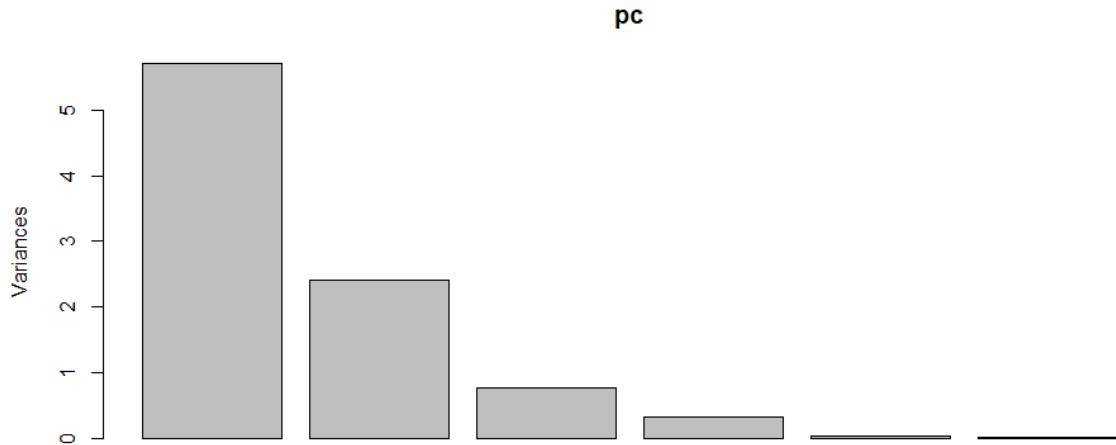
```
pc <- prcomp(subject) # 주성분분석 수행 함수
summary(pc) # 요약통계량
plot(pc)
```



```

> pc <- prcomp(subject)
> summary(pc)
Importance of components:
          PC1      PC2      PC3      PC4      PC5      PC6 
Standard deviation   2.389 1.5532 0.87727 0.56907 0.19315 0.12434 
Proportion of Variance 0.616 0.2603 0.08305 0.03495 0.00403 0.00167 
Cumulative Proportion 0.616 0.8763 0.95936 0.99431 0.99833 1.00000

```



1번째 PC1에는 61.6%의 분산비율을 차지하고 있다.

Proportion of Variance 0.616 0.2603 0.08305 0.03495 0.00403 0.00167

Cumulative Proportion 0.616 0.8763 0.95936 0.99431 0.99833 1.00000

$0.616 + 0.2603$  분산비율 = 0.8763 누적 비율(PC2)

$0.616 + 0.2603 + 0.08305 = 0.95936$  누적비율(PC3)

$0.616 + 0.2603 + 0.08305 + 0.03495 = 0.99431$  누적비율(PC4)

유사성을 가져도 3개까지 줄일 수 있지 않을까? 생각해볼 수 있다.

PC1: 자연과학 + 물리화학,

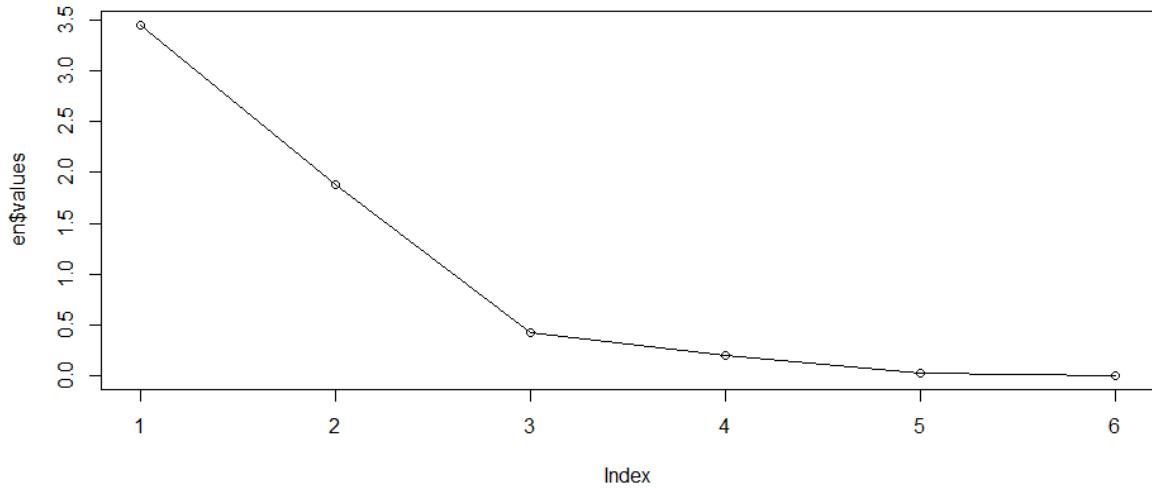
PC2: 인문사회 + 신문방송,

PC3: 응용수학 + 추론통계

대략적으로 90%가 넘어갈 때 3개까지 PC3 가져갈 수 있겠구나 판단할 수 있다.

3까지는 값의 변화가 급격하다.

4,5,6 값의 변화는 완만하다.



```
# 고유값으로 요인 수 분석
en <- eigen(cor(subject)) # $values : 고유값, $vectors : 고유벡터 # cor() 상관분석 알고리즘 # eigen() 고유값
names(en) # "values" "vectors"

en$values # $values : 고유값(스칼라) 보기
en$vectors
plot(en$values, type="o") # 고유값을 이용한 시각화
```

## 상관분석

### correlation

미국식[|kɔ:rə|lərfn; |ka:rə|lərfn] ↗ 영국식[|kɒrə|lərfn] ↗

(영사)

연관성, 상관관계

There is a direct correlation between exposure to sun and skin cancer. ↗

햇볕 노출과 피부암 사이에는 직접적인 연관성이 있다.

[영어사전 결과 더보기](#)

상관성은 **-1 ~ 1** 까지이다.

얼마나 나와 상관있는지 비교할 수 있다.

```
> en$vectors
[,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] -0.4962499 -0.351093036 0.63460534 0.3149622 0.45699508 0.03041553
[2,] -0.4319311 -0.400526644 0.11564711 -0.4422216 -0.57042232 0.34452594
[3,]  0.2542077 -0.628807884 -0.06984072 0.3339036 -0.35389906 -0.54622817
[4,]  0.3017115 -0.566028650 -0.37734321 -0.2468016 0.50326085 0.36333366
[5,]  0.4763815  0.008436692 0.58035475 -0.6016209 0.05643527 -0.26654314
[6,]  0.5155637  0.021286661 0.31595023 0.4133867 -0.28995329 0.61559319
```

상관성이 마이너스인 경우에는 상관성이 아주 강한 것이다, 다만 방향성은 다를 뿐.

예: 변비약1 ~ 설사약-1, 마이크와 물, 눈과 안경

하지만 0은 전혀 상관이 없다.

```
> # [실습] 변수 간의 상관관계 분석과 요인분석
> cor(subject)
      s1        s2        s3        s4        s5        s6
s1  1.00000000  0.86692145  0.05847768 -0.1595953 -0.5504588 -0.6262758
s2  0.86692145  1.00000000  0.06745441 -0.0240123 -0.6349581 -0.7968892
s3  0.05847768  0.06745441  1.00000000  0.9239433  0.3506967  0.4428759
s4 -0.15959528 -0.02401230  0.92394333  1.0000000  0.4207582  0.4399890
s5 -0.55045878 -0.63495808  0.35069667  0.4207582  1.0000000  0.8733514
s6 -0.62627585 -0.79688923  0.44287589  0.4399890  0.8733514  1.0000000
```

```
Uniquenesses:
      s1        s2        s3        s4        s5        s6
 0.005  0.056  0.051  0.005  0.240  0.005
```

6개의 유효성에 대한 것을 가지는 지 보여줌. (의미 O, 의미 X를 수치적으로 계산해서 보여줌.)

0.5 이하의 값이면 문제가 없다는 의미.

#### Loadings 요인 적재값.

```
Loadings:
  Factor1 Factor2 Factor3
s1   -0.379      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
```

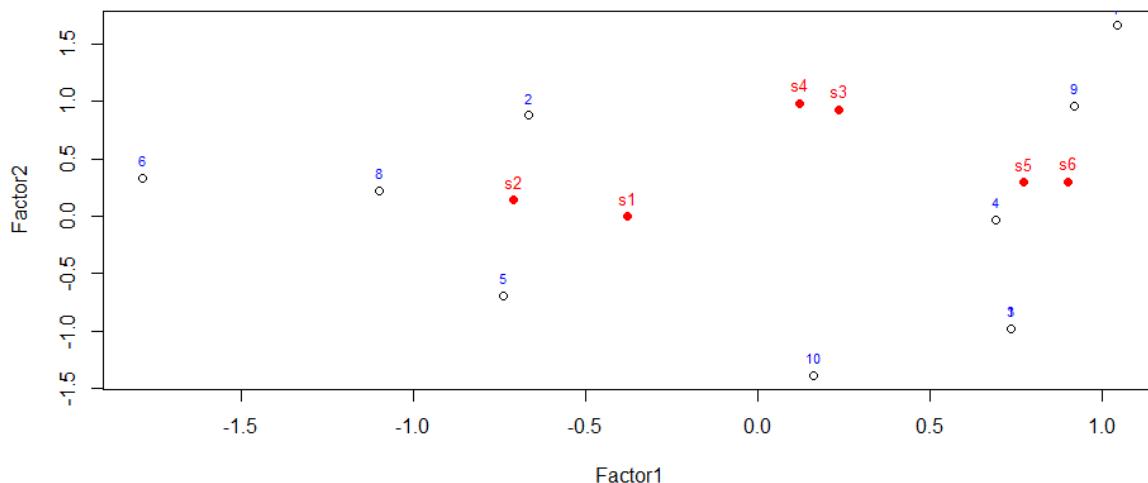
factor1과 변수 s5, s6(유사성이 깊다)을 한꺼번에 둑어도 된다고 제시해주는 것.

```
Factor1 Factor2 Factor3
SS loadings   2.122   2.031   1.486 # 각 요인별 적재되어진 것 제곱의 합
Proportion Var 0.354   0.339   0.248 # 분산의 비율
Cumulative Var 0.354   0.692   0.940 # 누적해서 보여줌
```

- SS loadings - Factor1가 2.122로 가장 높다.
- Cumulative Var - (1-0.940) = 0.06의 손실이 있다. 하지만 손실이 너무 커지면 누적분산은 좋지 못하다.

어쨌든 Factor별로 둑어서 보면 된다.

Factor1과 Factor2 요인점수 행렬



```
# 요인점수를 이용한 요인적재량 시각화

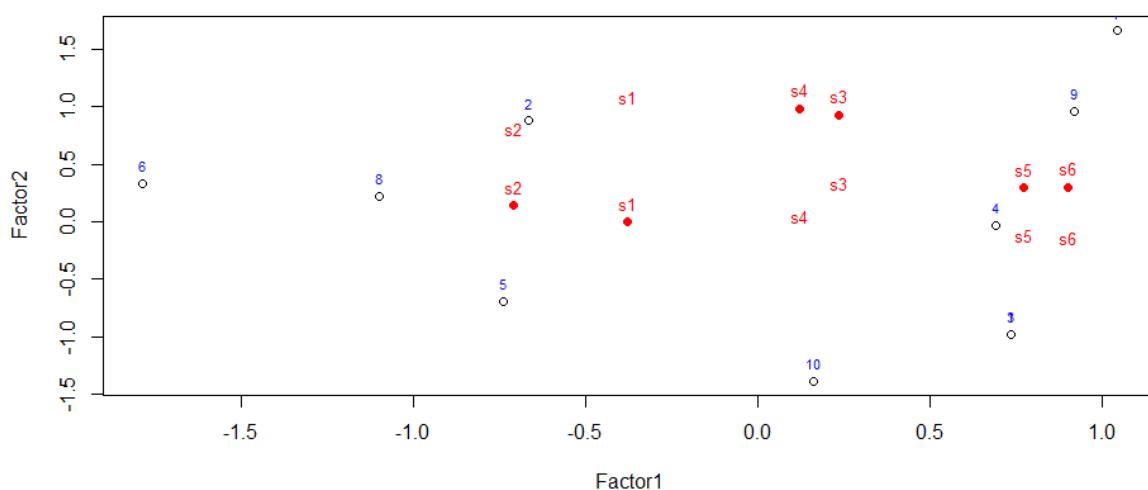
# (1) Factor1, Factor2 요인지표 시각화
plot(result$scores[, c(1:2)], main="Factor1과 Factor2 요인점수 행렬")
text(result$scores[, 1], result$scores[, 2],
    labels = name, cex = 0.7, pos = 3, col = "blue")

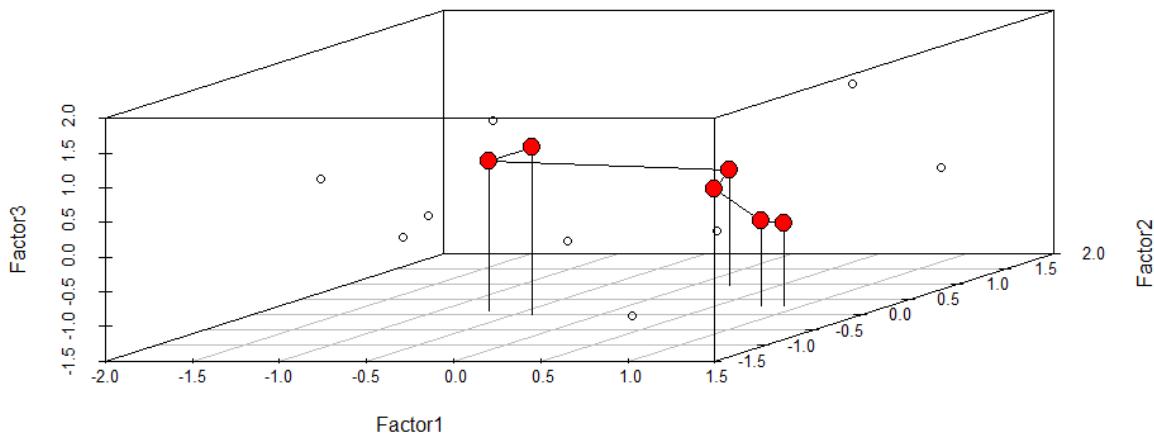
# 요인적재량 plotting
points(result$loadings[, c(1:2)], pch=19, col = "red")
text(result$loadings[, 1], result$loadings[, 2],
    labels = rownames(result$loadings),
    cex = 0.8, pos = 3, col = "red")

# (2) Factor1, Factor3 요인지표 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")

# 요인적재량 plotting
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")
```

Factor1과 Factor2 요인점수 행렬





```
# 3차원 산점도로 요인적재량 시각화
install.packages("scatterplot3d")
library(scatterplot3d)

Factor1 <- result$scores[,1]
Factor2 <- result$scores[,2]
Factor3 <- result$scores[,3]
# scatterplot3d(밀변, 오른쪽변, 원쪽변, type='p') # type='p' : 기본산점도 표시
d3 <- scatterplot3d(Factor1, Factor2, Factor3)

# 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')
```

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
	q8	음료의 복 담김
제품 만족도	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  $\pm 3$  이하) 그 자료는 요인 분석에 적합하지 않다.

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	

## 잘못 분류된 요인 제거로 변수 정제

Factor1: Q1,2,3,4

Factor2: Q1,2,3

Factor3: Q4,5,6,7

```
# 잘못 분류된 요인 제거로 변수 정제
Loadings:
  Factor1 Factor2 Factor3
Q1  0.201  0.762  0.240
Q2  0.172  0.813  0.266
Q3  0.141  0.762  0.340
Q4  0.250  0.281  0.641
Q5  0.162  0.488  0.557
Q6  0.224  0.312  0.693
Q7  0.235  0.219  0.703
Q8  0.695  0.225  0.304
Q9  0.873  0.122  0.155
Q10 0.852  0.144  0.161
Q11 0.719  0.152  0.225
```

# 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		

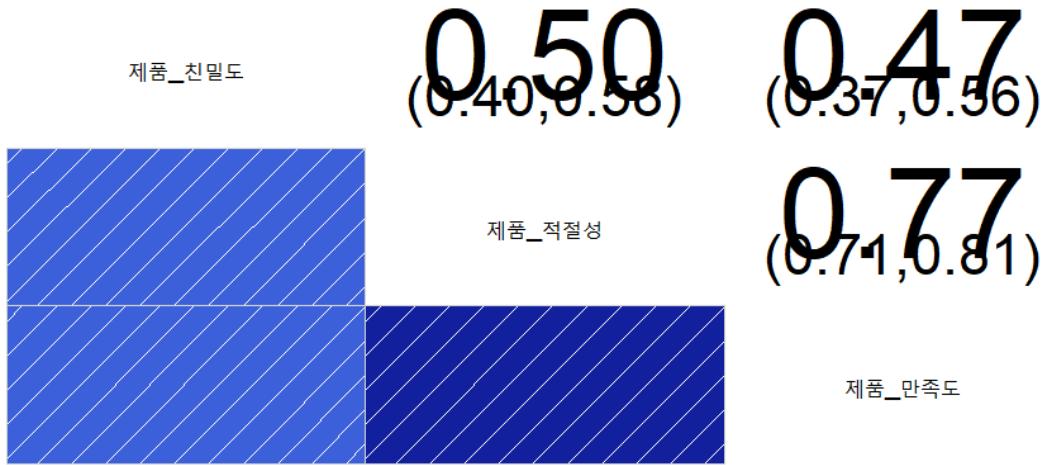
```
dw_df <- drinking_water_df[-4] #제외시킴
```

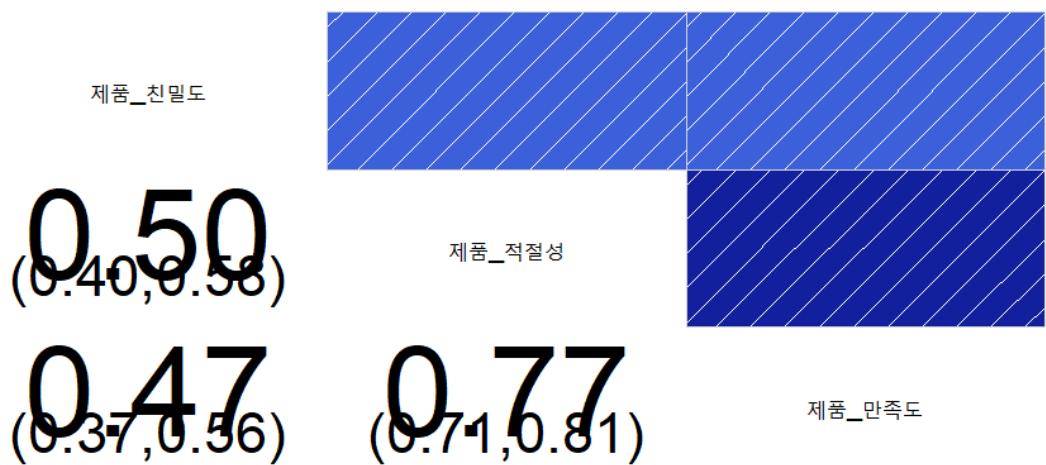
11개 → 3개로 차원 축소시킴(주성분분석)

```
> cor(drinking_water_factor_df)
  제품만족도 제품친밀도 제품적절성
제품만족도  1.0000000  0.4047543  0.4825335
제품친밀도  0.4047543  1.0000000  0.6344751
제품적절성  0.4825335  0.6344751  1.0000000
```



진할수록 숫자값이 높다.





```

## Chapter14-2. 상관관계 분석(Correlation Analysis)
#####
# 2.2 상관관계 분석 수행

# [실습] 기술 통계량 구하기
result <- read.csv("D:/heaven_dev/workspaces/R/data/product.csv", header=TRUE)
View(result)
head(result) # 친밀도 적절성 만족도(등간척도 - 5점 척도)

# 기술통계량
summary(result) # 요약통계량

sd(result$제품_친밀도); sd(result$제품_적절성); sd(result$제품_만족도)

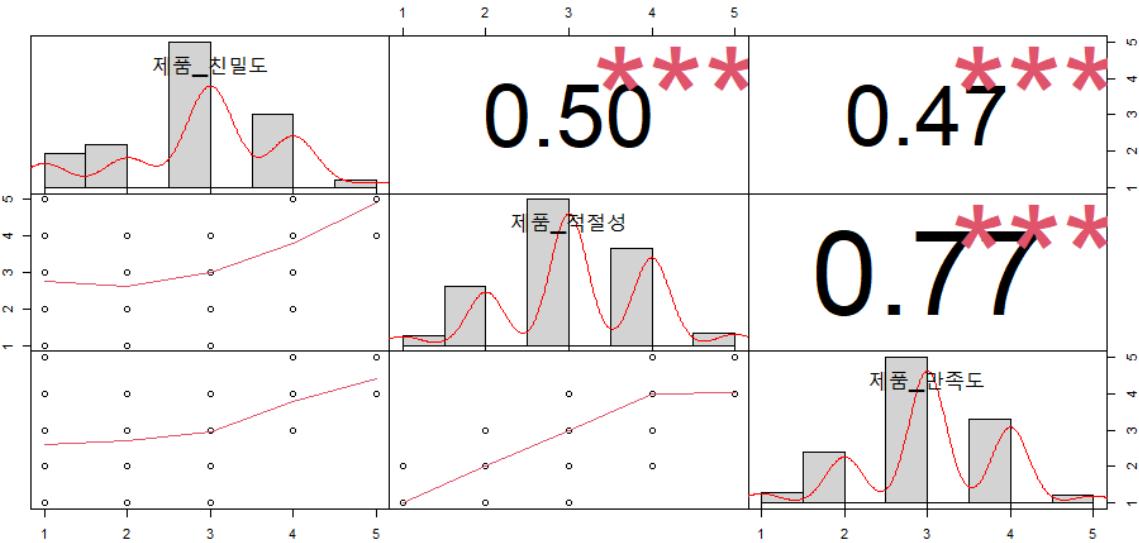
# [실습] 상관계수(coefficient of correlation) : 두 변수 X,Y 사이의 상관관계 정도를 나타내는 수치(계수) # 0.4 이상이 되면 상관도를 높게 봄.
cor(result$제품_친밀도, result$제품_적절성) # 0.4992086 -> 다소 높은 양의 상관관계
cor(result$제품_친밀도, result$제품_만족도) # 0.467145 -> 다소 높은 양의 상관관계
cor(result)

# [실습] 전체 변수 간 상관계수 보기
cor(result, method="pearson") # pearson - 데이터를 조사하면서 하나의 대표값을 찾아내는 개념

# [실습] 방향성 있는 색상으로 표현
install.packages("corrgram")
library(corrgram)
corrgram(result) # 색상 적용 - 동일 색상으로 그룹화 표시
corrgram(result, upper.panel=panel.conf) # 수치(상관계수) 추가(위쪽)
corrgram(result, lower.panel=panel.conf) # 수치(상관계수) 추가(아래쪽)

```

상관분석에 의한 시각화



```
# [실습] 방향성 있는 색상으로 표현
install.packages("corrgram")
library(corrgram)
corrgram(result) # 색상 적용 - 동일 색상으로 그룹화 표시
corrgram(result, upper.panel=panel.conf) # 수치(상관계수) 추가(위쪽)
corrgram(result, lower.panel=panel.conf) # 수치(상관계수) 추가(아래쪽)

# [실습] 차트에 밀도 곡선, 상관성, 유의학률(별표) 추가
install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)

# 상관성, p값(*), 정규분포 시각화 - 모수 검정 조건
chart.Correlation(result, histogram=, pch="+")

# [실습] spearman : 서열적도 대상 상관계수
cor(result, method="spearman")
```

```
> cor(result, method="spearman")
      제품_친밀도 제품_적절성 제품_만족도
제품_친밀도 1.0000000 0.5110776 0.5012007
제품_적절성 0.5110776 1.0000000 0.7485096
제품_만족도 0.5012007 0.7485096 1.0000000
```

## 시계열 - 미래에 대한 예측 분석

머신러닝, 딥러닝도 미래를 예측 추론 분석하고자 만듦.

