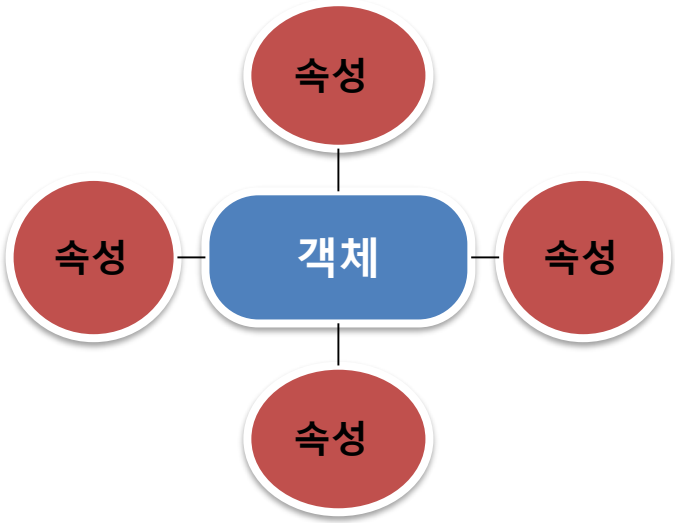
기술 통계 분석

변수(변인)

- 변수(Variable)
  - 변수(변인) 연구 대상 ▶ 객체(Object)
  - 분석되는 단위
  - 속성으로 구성
  - 예, **성별**(1=남자, 2=여자)



- 인구 통계학적 변수
  - 성장하면서 만들어지는 변수
  - 개인을 구별해 주는 속성
  - 성별, 연령, 학력, 종교, 생활수준 등

변수(변인)

- 변수의 유형
  - ① 독립변수(Independent variable) : 종속변수에 영향을 주는 변수
    예: 교육시간(독립)이 판매액(종속)에 영향을 미치는가?
  - ② 종속변수(dependent variable) : 독립변수의 영향을 받아 변화될 것으로 예측되는 변수
  - ③ 통제변수(Control variable) : 표본에 대한 일정한 수준의 값이 유지되게 하는 변수

[**가설**] 아이에게 모유를 먹이는 것이 어머니와 아이의 친근감과 따뜻함을 증가시킨다.
[**검정**] 모유를 먹이지 않은 어린이, 1~-5개월 먹인 어린이, 5개월 이상 먹인 어린이 들을 대상으로 2살 된 어린이들을 찾아 보았다. 이 어린이들과 어머니들을 2시간 동안 같이 지내게 하는 상황에서 어머니와 아이의 가까운 정도를 측정했다.

- 독립변수 : 모유 먹인 기간(예 : 3수준 척도 : ① 0 ② 1~5 ③ 5개월 이상)
- 종속변수 : 어린이와 어머니의 관계에 대한 가까운 정도
- 통제변수 : 2살 된 어린이

변수(변인)

- 변수의 유형

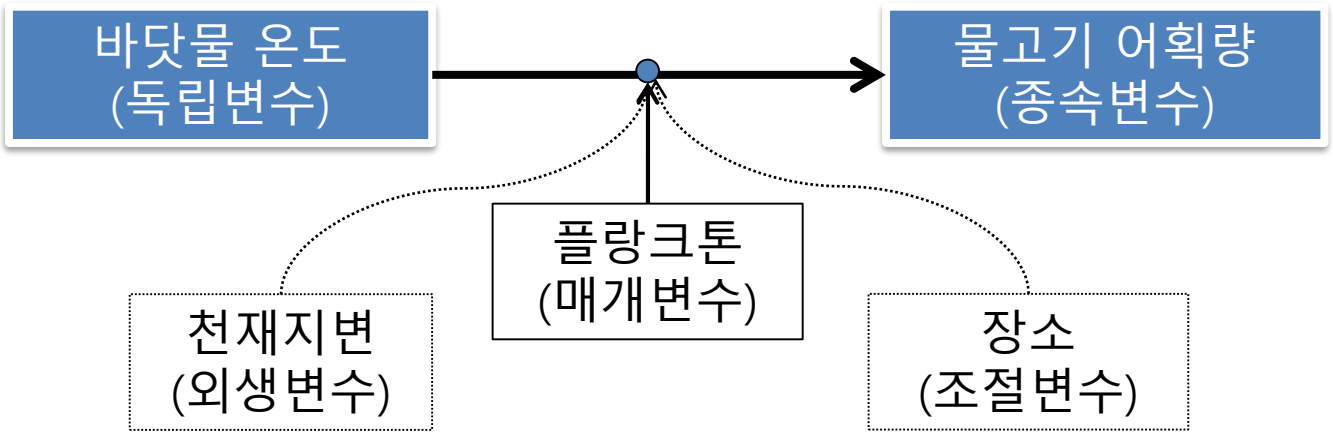
① 독립변수(Independent variable) : 종속변수에 영향을 주는 변수(설명)

② 종속변수(dependent variable) : 독립변수의 영향을 받아 변화될 것으로 예측되는 변수(성과, 반응)

③ 매개변수 : 두 변수를 중간에서 연결 시켜주는 변수

④ 조절변수 : 독립변수와 종속변수간 관계의 강도를 조절해주는 변수

⑤ 외생변수 : 독립변수와 종속변수의 관계를 잘못 이해하게 만드는 변수



척도

- 척도(Scale)
  - 변수에 값을 부여하는 방법
  - 변수 측정 단위(응답자가 선택할 수 있는 질문 항목)

정성적-질적 척도(범주형 변수)		정량적-양적 척도(연속형 변수)	
명목척도	이름이나 범주를 대표하는 의미 없는 숫자 (예 : ① 남자 ② 여자)	**등간척도**	속성에 대한 각 수준 간의 간격이 동일한 경우(가감산 연산) (예: 연소득이 어디에 해당되십니까?)
서열척도	측정 대상 간의 높고 낮음(서열), 순서에 대한 값 부여 (예 : 좋아하는 순위를 표시하시오.)	**비율척도**	등간척도의 특성에 절대원점(0)이 존재하고, 비율계산이 가능한 경우(사칙연산) (예 : 나이가 몇 세 입니까?)

척도

- 명목척도(Nominal scale)
  - 단순히 속성을 분류할 목적으로 명목상 숫자를 부여한 척도
  - 연산 불가능한 변수(연산은 가능하지만 의미가 없다.)

❖ 예) 성별(1=남자, 2=여자), 연령별, 학력, 종교, 취미, 선수번호 등

설문지 예문) 본인의 최종학력을 표시하시오.

① 초졸 ② 중졸 ③ 고졸 ④ 대졸 ⑤ 대학원졸

척도

- 서열척도(Ordinal scale)
  - 측정대상 간의 크고 작음, 양의 많고 적음, 선호도의 높고 낮음
  - 순서관계를 밝혀주는 척도(연산 불가능한 변수)

❖ 예) 시험 성적에 대한 순위 관계, 키 순서 등

설문지 예문) 가장 좋아하는 음료수의 순서대로 1,2,3,4의 숫자를 표시하시오.

커피( )　　녹차( )　　홍차( )　　우유( )

척도

- 등간척도(Interval scale)
  - 측정대상의 속성에 대한 각 수준 간의 간격이 동일한 척도
  - 덧셈과 뺄심 연산 가능 변수(배수 관계 없음)
  - 절대원점(0)을 가지고 있지 않음(의미 없음)
  - 설문지 작성에서 가장 많이 이용
  - 시각(년도, 시각, 월), 섭씨온도, 화씨온도

설문지 예문) 연수 교재는 학생상담에 유용한 자료가 되었습니까? (5점 척도)

① 전혀그렇지 않다. ② 그렇지않다. ③ 보통이다. ④ 그렇다. ⑤ 매우그렇다.

척도

- 비율척도(Ratio scale)
  - 척도의 수가 등간
  - 절대원점(0)을 가지고 있는 척도(0을 기준으로 한 수치)
  - 사칙연산 모두 가능
  - 등간척도와 함께 많이 사용되는 변수
  - 예) 성적, 키, 무게, 인구수, 수량, 길이, 금액 등

설문지 예문) 귀하의 몸무게는 얼마입니까?

( )kg

척도

- 통계분석 방법과 변수척도 관계

분석방법	적용분야	변수척도
빈도분석	가장 기초적이고 간단한 분석방법	**모든 척도**
교차분석 (카이제곱)	변수 간의 교차표 작성	명목척도, 서열척도
요인분석	• 타당성 검정 • 설명력 부족한 변수 제거	**등간척도,비율척도**
신뢰도분석	추출된 요인들의 동질적인 변수 구성	**등간척도,비율척도**
상관관계분석	측정변수들 간의 관계 정도를 제시	**피어슨 - 등간척도, 비율척도** 스피어만 - 서열척도
회귀분석	인과관계 분석	**독립변수, 종속변수 : 등간척도/비율척도**
t-검정	집단 간 평균 차이 검정	**독립변수 :** 명목척도 **종속변수 : 등간척도 또는 비율척도**
분산분석 (ANOVA)	3집단 이상의 평균 검정	**독립변수 :** 명목척도 **종속변수 : 등간척도 또는 비율척도**

척도별 기술통계

- 기술통계 (Descriptive Statistics)

  - 자료를 요약하는 기초적인 통계량
  - 데이터 분석 전에 전체적인 데이터 분포의 이해
  - 데이터의 분석 방향 고려
  - 기술통계량을 통해서 모집단 특성 유추

척도별 기술통계

- 척도 유형

resident	gender	age	level	cost	type	survey	pass
거주지역	성별	나이	학력수준	생활비	학교유형	만족도	합격여부
명목	명목	비율	서열	비율	명목	등간	명목
1~5	1, 2	40~69	1, 2, 3	2~10	1, 2	1~5	1, 2

척도별 기술통계

1) 척도별 기술통계량

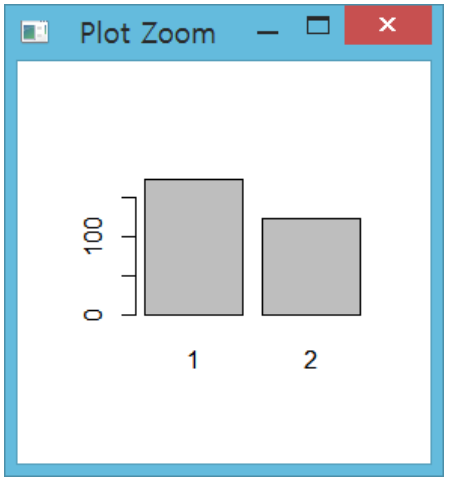
- 데이터 특성 보기(전체 데이터 대상)

```
dim(data) # 행(300)과 열(8) 정보 - 차원보기
length(data) # 열(8) 길이
length(data$survey) #survey 컬럼의 관찰치 - 행(300)

str(data) # 데이터 구조보기 -> 데이터 종류,행/열,data
# 'data.frame':          300 obs. of  8 variables:
str(data$survey) # int [1:300] 1 2 1 4 3 3 NA NA NA 1 ...

# 데이터 특성(최소, 최대, 평균, 분위수, 노이즈-NA) 제공
summary(data)
```

척도별 기술통계

- 명목척도 변수의 기술통계량

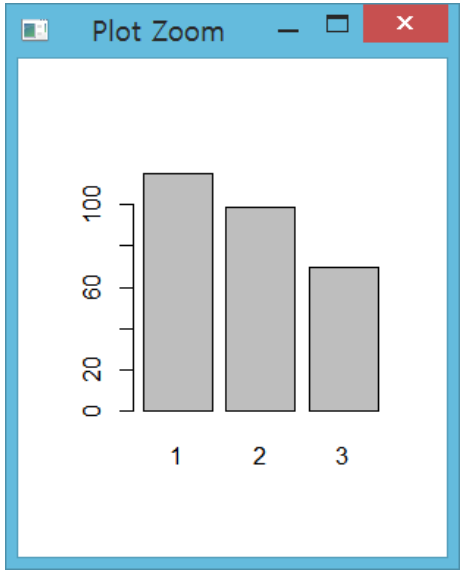
```
# 명목상 의미 없는 수치로 표현된 변수 - 성별(gender)
length(data$gender)
summary(data$gender) # 최소, 최대, 중위수, 평균 – 의미 없음
table(data$gender) # 각 성별 빈도수 - outline 확인-> 0, 5

# 성별 outline제거
data <- subset(data,data$gender == 1 | data$gender == 2)
# data 테이블을 대상으로 성별이 1 또는 2인 데이터 대상 subset 구성
barplot(x) # 범주형(명목/서열척도) 시각화 -> 막대차트

prop.table(x) # 비율 계산 : 0< x <1 사이의 값
y <-  prop.table(x)
round(y*100, 2) #백분율 적용(소수점 2자리)
# 1:58.25, 2:41.75
```



척도별 기술통계

- 서열척도 변수의 기술통계량

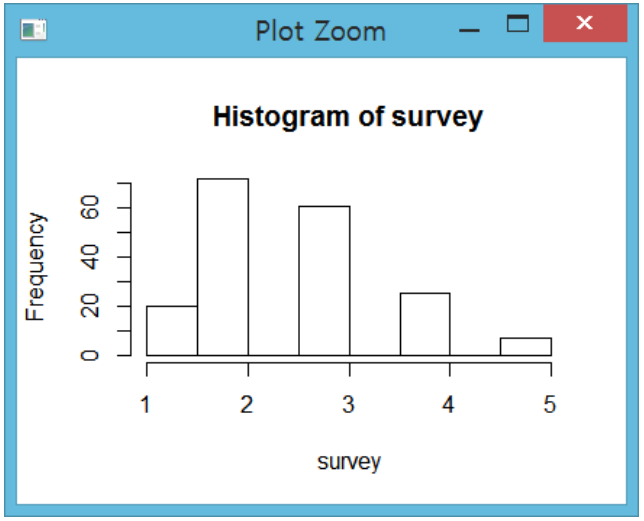
```
# 계급순위를 수치로 표현한 변수 - 학력수준(level)
length(data$level) # 학력수준 - 서열
summary(data$level) # 명목척도와 함께 의미없음
table(data$level)  # 빈도분석 - 의미있음

x1 <- table(data$level) # 각 학력수준에 빈도수 저장
x1
barplot(x1) # 명목/서열척도 -> 막대차트
#  1   2   3
# 115  99  70  <- 빈도분석 결과
```



척도별 기술통계

- 등간척도 변수의 기술통계량

```
# 속성의 간격이 일정한 변수(survey) - 덧셈/뺄셈 연산 가능
survey <- data$survey
survey
summary(survey) # 만족도(5점 척도)인 경우 의미 있음 -> 2.6(평균이상)
x1<-table(survey) # 빈도수
x1
#1  2  3  4  5
#20 72 61 25  7
hist(survey)
# 연속형 척도 시각화 -> 범주화 -> 히스토그램
```



척도별 기술통계

- 비율척도 변수의 기술통계량

```
# 수치로 직접 입력한 변수(cost)
length(data$cost)
summary(data$cost) # 요약통계량 – 의미 있음(mean) - 8.784
mean(data$cost) # NA
data$cost

# 데이터 정제 - 결측치 제거 및 outline 제거
plot(data$cost)
data <- subset(data,data$cost >= 2 & data$cost <= 10) # 총점기준
data
x<- data$cost
x
mean(x) # 평균 : 5.354
# 평균이 극단치에 영향을 받는 경우 - 중위수(median) 대체
median(x) # 5.4
```

척도별 기술통계

```
min(x)
max(x)
range(x) # min ~ max
sort(x) # 오름차순
sort(x, decreasing=T) # 내림차순

sd(x) # 표준편차 - 1.138783
var(x) # 분산 - 1.296826
# 표준편차 : 표본의 평균에서 얼마나 떨어져 있는가 – 산포도

quantile(x, 1/4) # 1 사분위수 - 25%, 4.6
quantile(x, 3/4) # 3 사분위수 - 75%, 6.2
```

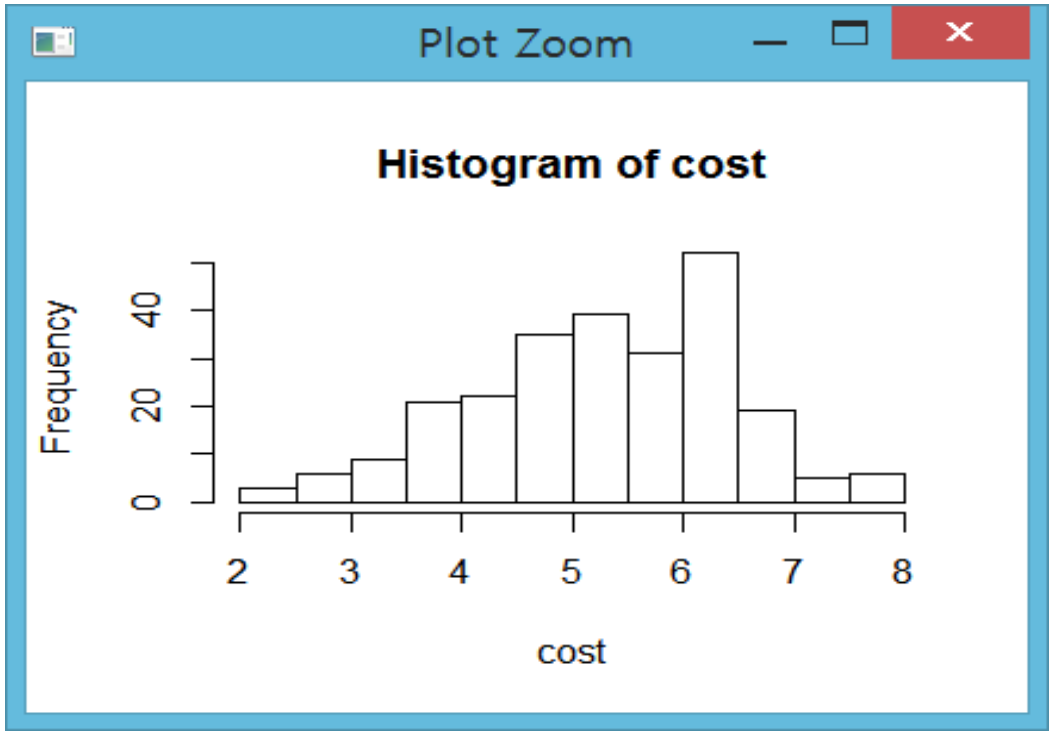
척도별 기술통계

- 패키지를 이용한 비대칭도 나타내기

```
install.packages("moments")  # 왜도/첨도 사용을 위한 패키지 설치
library(moments)
cost <- data$cost      kp
# 왜도 – 평균 중심으로 기울어짐 정도
skewness(cost) # -0.2974908
# 0보다 작으면, 왼쪽방향 비대칭 꼬리, 0보다 크면, 오른쪽 방향 비대칭 꼬리,
# 0에 근사하면 중심으로 좌우대칭

#첨도 - 표준정규분포와 비교하여 얼마나 뽀족한가 측정 지표
kurtosis(cost) # 2.683438

# 표준정규분포와 비교하여 첨도가 3이며 정규분포 곡선을 이루고,
# 첨도가 3보다 크면 정규분포 보다 뽀족한 형태, 3보다 작으면
# 정규분포 보다 완만한 형태이다.

hist(cost) # 히스토그램으로 왜도/첨도 확인
# 왼쪽방향 비대칭 꼬리, 정규분포 첨도 보다 완만함
```

척도별 기술통계

- 왜도/첨도에 의한 비대칭도 시각화



❖ 데이터가 정규분포 형태를 띄고 있는가의 여부를 알기 위해서 비대칭도를 이용한다.

척도별 기술통계

2) 패키지 이용 기술통계량 구하기

- Hmisc 패키지 이용

```
install.packages("Hmisc") #  패키지 설치
library(Hmisc) #  패키지 메모리 로딩

# 전체 변수 대상 기술통계량 제공 - 빈도와 비율 데이터 일괄 수행
describe(data) # Hmisc 패키지에서 제공되는 함수
# 명목,서열,등간척도 - n, missing,unique, 빈도수,비율
# 비율척도 - n, missing, unique, mean, lowest, highest

# 개별 변수 기술통계량
describe(data$gender) # 특정 변수(명목) 기술통계량 - 비율 제공
describe(data$age) # 특정 변수(비율) 기술통계량 - lowest, highest

summary(data$age)
```

척도별 기술통계

- prettyR 패키지 이용

```
# Hmisc 패키지 보다 유용
install.packages("prettyR")
library(prettyR)

# 전체 변수 대상
freq(data) # 각 변수별 : 빈도, 결측치, 백분율, 특징-소수점 제공
# 개별 변수 대상
freq(data$gender) # 빈도와 비율 제공
```

척도별 기술통계

3) 기술통계량 보고서 데이터 작성

```
 # 거주지역 변수 리코딩
data$resident2[data$resident == 1] <-"특별시"
data$resident2[data$resident >=2 & data$resident <=4] <-"광역시"
data$resident2[data$resident == 5] <-"시구군"

x<- table(data$resident2)
prop.table(x) # 비율 계산 : 0< x <1 사이의 값
y <-  prop.table(x)
round(y*100, 2) #백분율 적용(소수점 2자리)
#광역시 시구군 특별시
#37.66  14.72  47.62
```

척도별 기술통계

```
 # 성별 변수 리코딩
data$gender2[data$gender== 1] <-"남자"
data$gender2[data$gender== 2] <-"여자"

x<- table(data$gender2)
prop.table(x) # 비율 계산 : 0< x <1 사이의 값
y <-  prop.table(x)
round(y*100, 2) #백분율 적용(소수점 2자리)
#남자  여자
#58.87 41.13
```

척도별 기술통계

```
# 나이 변수 리코딩
data$age2[data$age <= 45] <-"중년층"
data$age2[data$age >=46 & data$age <=59] <-"장년층"
data$age2[data$age >= 60] <-"노년층”

x<- table(data$age2)
prop.table(x) # 비율 계산 : 0< x <1 사이의 값
y <-  prop.table(x)
round(y*100, 2) #백분율 적용(소수점 2자리)
#노년층 장년층 중년층
#24.60  68.15   7.26
```

척도별 기술통계

```
# 학력수준
data$level2[data$level== 1] <-"고졸"
data$level2[data$level== 2] <-"대졸"
data$level2[data$level== 3] <-"대학원졸"

x<- table(data$level2)
prop.table(x) # 비율 계산 : 0< x <1 사이의 값
y <-  prop.table(x)
round(y*100, 2) #백분율 적용(소수점 2자리)
#고졸      대졸 대학원졸
#39.41     36.44     24.15
```

척도별 기술통계

```
# 합격여부 리코딩
data$pass2[data$pass== 1] <-"합격"
data$pass2[data$pass== 2] <-"실패"

y<- table(data$pass2)
prop.table(x) # 비율 계산 : 0< x <1 사이의 값
y <-  prop.table(x)
round(y*100, 2) #백분율 적용(소수점 2자리)
#고졸      대졸 대학원졸
#39.41     36.44     24.15

  head(data)
```

resident	gender	age	level	cost	type	survey	pass	cost2	resident2	gender2	age2	level2	pass2
1	1	1	50	1	5.1	1	1	2	2	특별시	남자	장년층	고졸	실패
2	2	1	54	2	4.2	1	2	2	2	광역시	남자	장년층	대졸	실패
3	NA	1	62	2	4.7	1	1	1	2	<NA>	남자	노년층	대졸	합격
4	4	2	50	NA	3.5	1	4	1	NA	광역시	여자	장년층	<NA>	합격
5	5	1	51	1	5.0	1	3	1	2	시구군	남자	장년층	고졸	합격
6	3	1	55	2	5.4	1	3	NA	2	광역시	남자	장년층	대졸	<NA>

기술통계량 보고서

❖논문에서 응답자의 인구 통계적 특성은 반드시 제시 하여야 한다.

<인구통계적 특성 결과 제시>----------

'부모의 생활수준과 자녀의 대학진학 여부와 관련성이 있다.'를 분석하기 위해서 자녀를 둔 A회사 225명의 부모를 대상으로 거주지, 성별, 나이, 학력수준, 진학여부 등의 항목을 설문으로 조사하고, 정제된 데이터를 토대로 빈도분석을 실시하였다. 분석결과 전체 응답자 중에서 부모의 학력수준은 고졸이 93명으로 39.41%를 차지하여 가장 높은 빈도수를 나타냈고, 자녀의 성별 비율은 남자가 146명으로 58.87%를 차지하고, 여학생은 102명으로 41.13%을 차지하였다. 또한 자녀의 대학진학여부에서 합격은 139명으로 59.15%를 차지하고, 실패는 96명으로 40.85%를 차지한 것으로 나타났다.

기술통계량 보고서

표본의 인구통계적 특성 결과

변수		빈도수	구성비율(%)
거주지	특별시 광역시 시구군	89 34 111	38.03 14.53 47.44
성별	남자 여자	146 102	58.87 41.13
나이	장년층 중년층 노년층	172 18 61	68.53 7.17 24.30
학력수준	고졸 대졸 대학원졸	95 87 57	39.75 36.40 23.85
진학여부	실패 성공	98 140	41.18 58.82