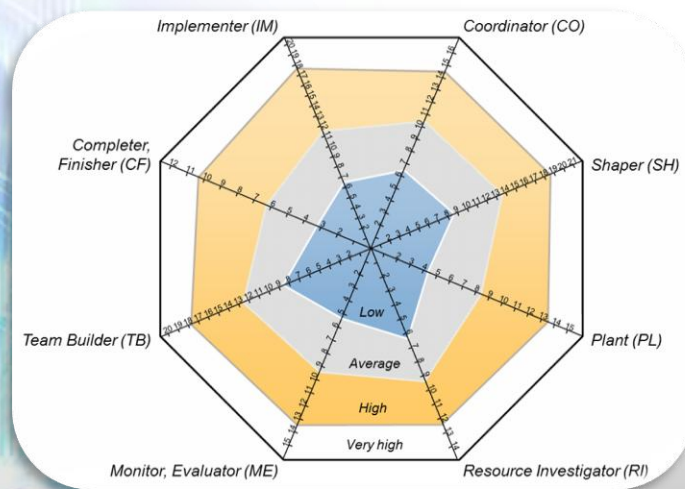


## Part-III. 추론통계 분석



- 10. 분석절차와 통계지식
- 11. 기술통계 분석
- 12. 교차분석과 Chi-square 분석
- 13. 집단 간 차이 분석
- 14. 요인분석과 상관분석

## 10-1. 통계분석 절차



단계0. 연구조사

단계1. 가설설정

단계2. 유의수준 결정

단계3. 측정도구 선정

단계4. 데이터 수집

단계5. 데이터 코딩

단계6. 통계분석 수행

단계7. 결과분석



# I. 통계분석 절차

## ● 논문/보고서 작성을 위한 통계분석 절차

1

가설설정

2

유의수준 결정

3

측정도구 선정

4

데이터 수집(설문지, 웹, SNS)

5

데이터 코딩/프로그래밍

6

통계분석 수행(R, SPSS, SAS)

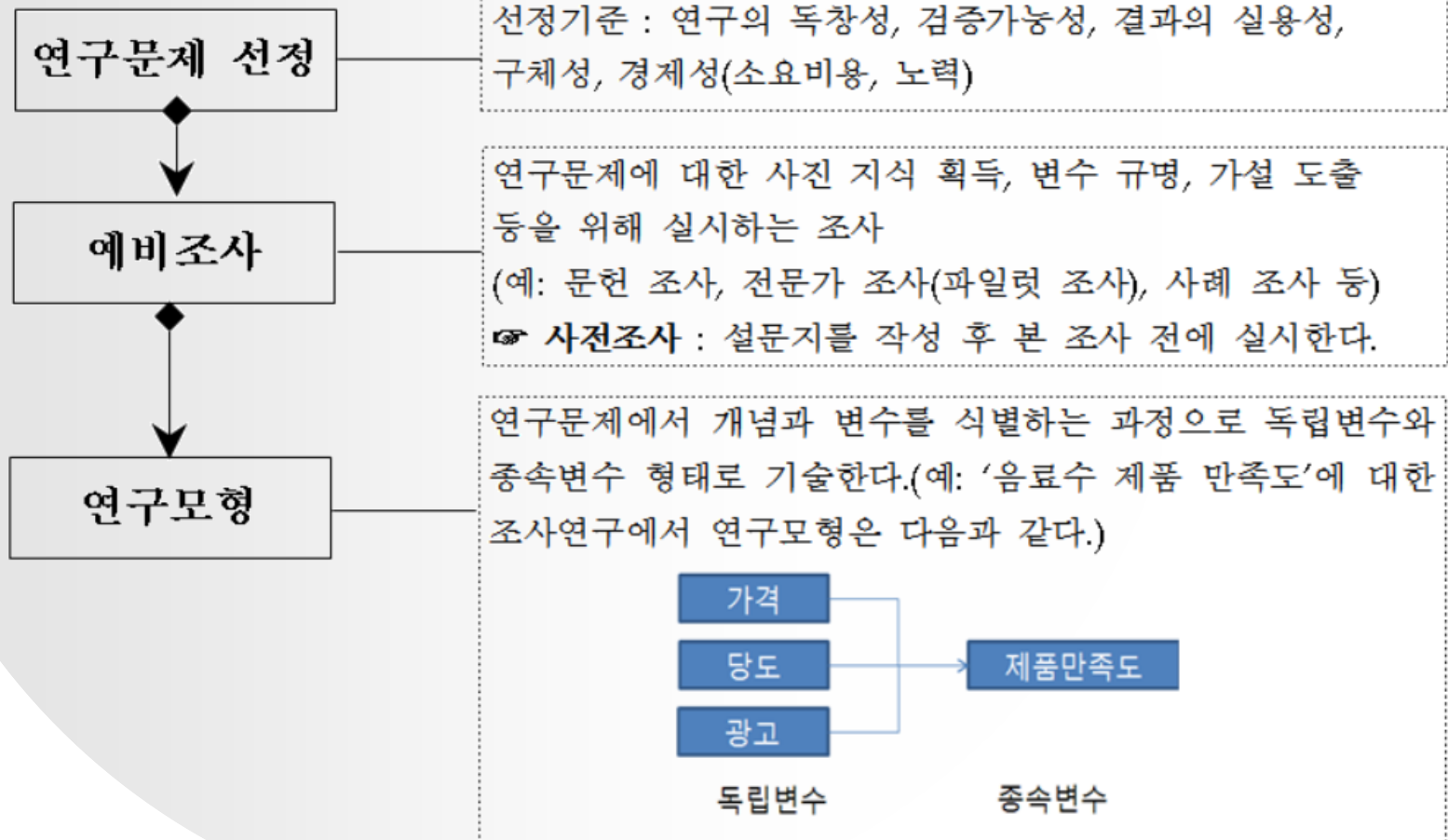
7

결과분석(논문/보고서 작성)



# 단계0. 연구조사

## ● 가설 설정 이전의 연구조사





# 단계1. 가설 설정

## ● 가설(Hypothesis)

- 사회 조사.연구에서 주어진 연구 문제에 대한 예측적 해답
- 실증적인 증명에 앞서 세우는 잠정적인 진술
- 나중에 논리적으로 검정될 수 있는 명제
- 통계분석을 통해서 채택 또는 기각

※ 과학적 연구에서 가설의 설정은 매우 중요



# 단계1. 가설 설정

## ● 가설의 유형

### ① 귀무가설(영가설)

'두 변수간의 관계가 없다.' 또는 '차이가 없다.'

- ✓ 부정적 형태 진술(예,  $H_0$  : 교육수준에 따라서 사회 정책에 대한 비판적 태도에서 차이가 없다.)

### ② 연구가설(대립가설)

'차이가 있다.' 또는 '효과가 있다.'

- ✓ 긍정적 형태 진술(예,  $H_1$ : 영양소별 효과의 차이는 있다.)

※ 논문에서 **연구가설 제시**, 귀무가설을 통해서 가설 검정





## 단계2. 유의수준과 임계값 결정(1/3)

$H_1$  = '신약A는 A암 치료에 효과가 있다.'

$H_0$  = '신약A는 A암 치료에 효과가 없다.'

- 분석결과 : 생쥐 100마리를 대상으로 신약A를 투약한 결과 검정통계량의 유의확률( $P=0.03$ )이 나왔다.
  - 이때 귀무가설은 기각되는가? ➔ **YES**
- 사회과학분야 임계값 :  $\alpha=0.05$ ( $p<0.05$ (5%미만))
  - 적어도 96마리 이상 효과 ➔  **$H_1$  채택**
- 의.생명분야 임계값 :  $\alpha=0.01$ (99% 신뢰도 보장)
  - 적어도 99마리 이상 효과 ➔  **$H_1$  채택**



## 단계2. 유의수준과 임계값 결정(2/3)

### ● 유의수준(Significant level)

- 가설 채택 또는 기각 기준
- 분석 결과 유의수준 이내 → 가설 채택(그렇지 않으면 기각)
- $\alpha$ (알파) 표시
- 유의수준의 임계값(기준값) 결정
  - ✓ 일반 사회과학분야 :  $\alpha=0.05$ ( $p<0.05$ ) 기준
  - ✓  $\alpha=0.05$  : 통계치가 모수치를 대표하는 허용 오차 5%(신뢰도 95%)  
(예, 100번 가운데 5번 미만 나올 확률)
- 의생명분야 : 오차범위 최소  $\alpha=0.01$ (1% 오차 허용, 99% 신뢰도 확보)





## 단계2. 유의수준과 임계값 결정(3/3)

### ● 유의수준 $\alpha$ 와 P값 관계

$\alpha > P\text{값}$  : 연구가설 채택(귀무가설 기각)

$\alpha \leq P\text{값}$  : 연구가설 기각(귀무가설 채택)

단정적  
표현 不

- 귀무가설( $H_0$ ) : '영양소별 효과의 차이는 없다'에서 임계값( $\alpha=0.05$ ) 일때 가설 검정 결과 확률(p값) 0.04가 나왔다면  $p(0.04) < \alpha(0.05) \rightarrow$  귀무가설(영가설) 기각
- 영양소별 효과의 차이가 있을 확률이 높기 때문에 연구가설 채택
- 이때 통계적으로 유의하다라고 해석,  $p < 0.01$ 이면 매우 유의하다.  $p < 0.05$  수준이면 통계적으로 유의적인 차이를 보인다. '귀무가설이 의심스럽다'는 의미



## 단계3. 측정도구 선정

### ● 측정도구 선정

- 가설에 나오는 변수를 무엇으로 측정할 것인가를 결정하는 단계
  - 가설에 나오는 변수(변인) 추출
  - 변수의 척도를 고려 측정도구 선정
- ▶ 【척도(Scale)】 참조



## 단계4. 데이터 수집

### ● 데이터 수집(설문지 작성)

- 선정된 측정도구를 이용하여 설문 문항 작성 단계
- 조사응답자 대상 설문 실시 & 회수
- 정형/비정형 데이터 수집(DB, WEB, SNS 등 )
- 본 단계까지 완료된 경우
  - ✓ 연구목적과 배경, 연구모형, 연구가설까지 끝난 상태
  - ➔ 논문 50% 이상 완성



# 단계5. 데이터 코딩

## ● 데이터(설문지) 코딩

- 통계분석 프로그램(Excel, R, SPSS, SAS,) 데이터 입력
- 데이터 전처리(미 응답자, 잘못된 데이터 처리)

The screenshot shows a Microsoft Excel spreadsheet titled 'cleanDescriptive - Microsoft Excel'. The data is organized in columns A through N. The first row (row 1) contains headers: resident, gender, age, level, cost, type, survey, pass, cost2, resident2, gender2, age2, level2, pass2. The subsequent rows (rows 2-12) contain data entries. The data includes numerical values, categorical labels like '특별시', '광역시', '시구군', '남자', '여자', '장년층', '노년층', and 'NA', and a final column 'pass2' with values like '실패', '대출', '합격', and 'NA'.

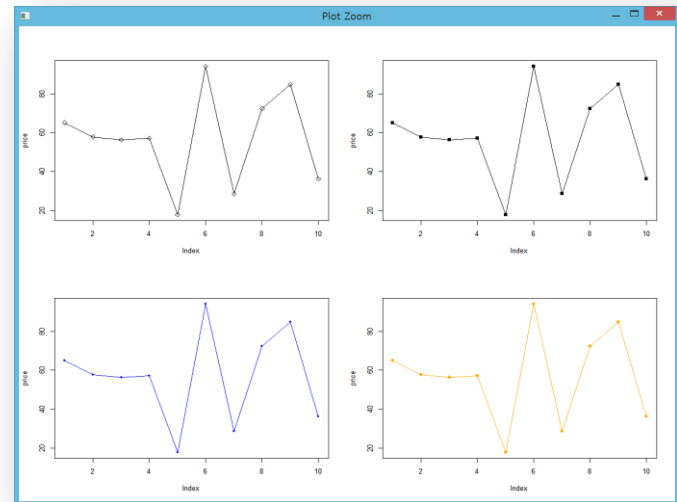
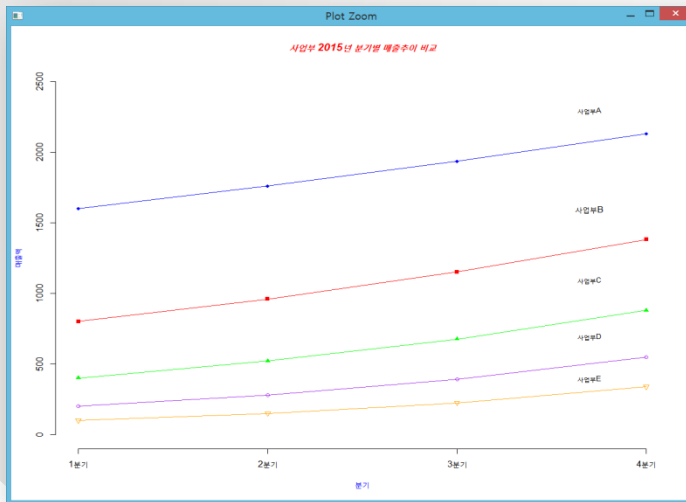
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	resident	gender	age	level	cost	type	survey	pass	cost2	resident2	gender2	age2	level2	pass2
2	1	1	50	1	5.1	1	1	2	2	특별시	남자	장년층	고출	실패
3	2	1	54	2	4.2	1	2	2	2	광역시	남자	장년층	대출	실패
4	NA	1	62	2	4.7	1	1	1	2	NA	남자	노년층	대출	합격
5	4	2	50	NA	3.5	1	4	1	NA	광역시	여자	장년층	NA	합격
6	5	1	51	1	5	1	3	1	2	시구군	남자	장년층	고출	합격
7	3	1	55	2	5.4	1	3	NA	2	광역시	남자	장년층	대출	NA
8	2	2	56	1	4.1	1	NA	2	2	광역시	여자	장년층	고출	실패
9	NA	1	49	1	4.4	1	NA	2	2	NA	남자	장년층	고출	실패
10	2	1	49	2	4.9	1	1	1	2	광역시	남자	장년층	대출	합격
11	5	2	49	NA	2.3	1	2	1	1	시구군	여자	장년층	NA	합격
12	3	1	52	1	4.2	1	2	2	2	광역시	남자	장년층	고출	실패



# 단계6. 통계분석 수행

## ● 통계분석 수행

- 전문 통계분석 프로그램(R, SPSS, SAS) 분석 단계
- ❖ 통계분석 방법을 계획하지 않고 데이터를 수집할 경우 실패 확률 높음



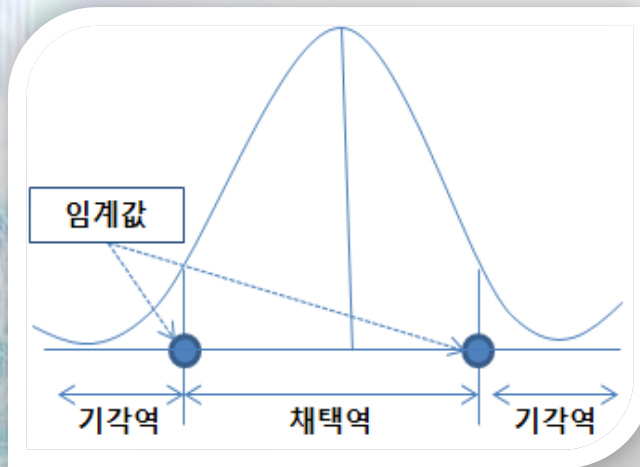


# 단계7. 결과분석

## ● 결과분석 제시

- 연구목적과 연구가설에 대한 분석 및 검증 단계
- 인구통계학적 특성 반영
- 주요 변인에 대한 기술통계량 제시
- 연구가설에 대한 통계량 검정 및 해석
- 연구자 의견 기술(논문/보고서 작성)

## 10-2. 통계 사전 지식



- 1) 통계학 개요
- 2) 모집단과 표본
- 3) 추정과 검정
- 4) 가설검정 오류
- 5) 검정통계량
- 6) 정규분포
- 7) 모수 & 비모수





# 1) 통계학 개요

## ● 통계학(Statistics)?

- ✓ 논리적 사고와 객관적인 사실에 의거, 확률 기반 인과관계 규명
- ✓ 특히 연구목적에 의해 설정된 가설들에 대하여 분석결과가 어떤 결과를 뒷받침하고 있는지를 통계적 방법으로 검정.
- ✓ 사회학, 경제학, 경영학, 정치학, 교육학, 공학, 의.생명 등 대부분의 모든 학문 분야에서 폭넓게 이용

구분	기술(Descriptive) 통계학	추론(Inferential) 통계학
기능	<ul style="list-style-type: none"><li>수집된 자료의 특성을 쉽게 파악하기 위해서 자료를 정리 및 요약</li></ul>	<ul style="list-style-type: none"><li>모집단에서 추출한 표본의 정보를 이용하여 모집단의 다양한 특성을 과학적으로 추론</li></ul>
방법	<ul style="list-style-type: none"><li>표, 그래프, 대푯값 등</li></ul>	<ul style="list-style-type: none"><li>회귀분석, T-검정, 분산분석 등</li></ul>



## 2. 모집단과 표본

### ① 전수조사

- 모집단내에 있는 모든 대상 조사 방법(예, 인구조사)
- 모집단의 특성 정확히 반영
- 시간과 비용이 많이 소요되는 단점

### ② 표본조사

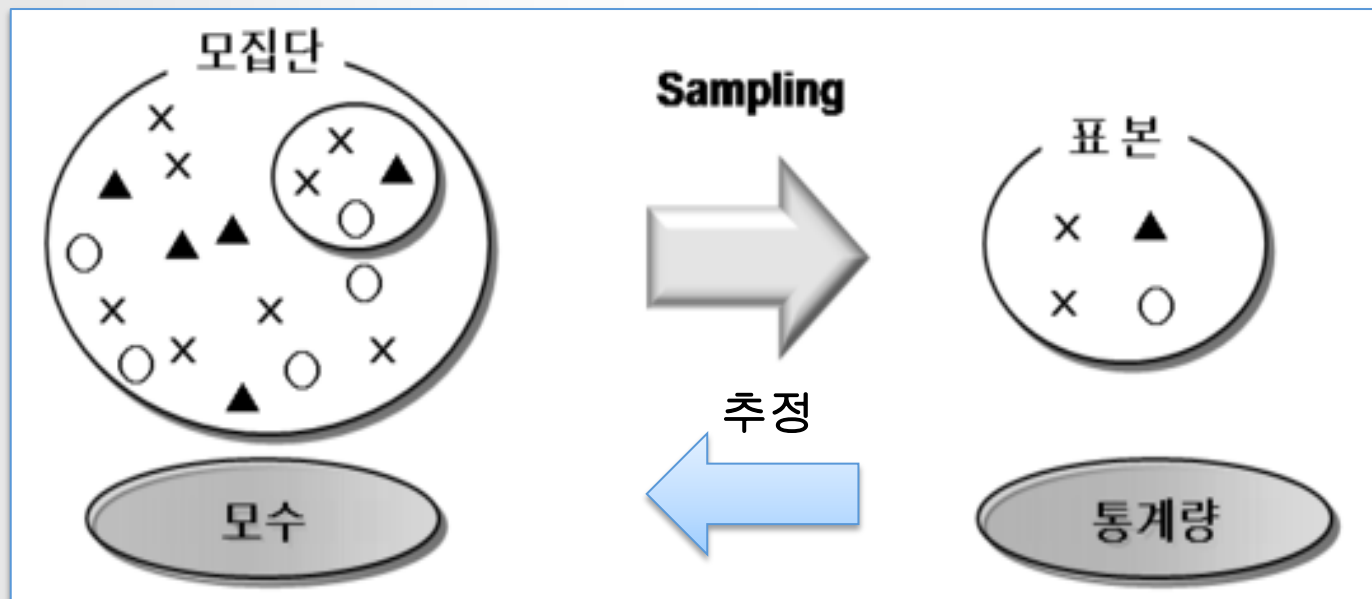
- 모집단으로부터 추출된 표본을 대상으로 분석 실시  
(예, 선거 여론조사, 마케팅조사, 안전성 검사, 의생명 임상실험)
- 모집단의 특성을 반영하지 못하는 표본은 무용지물



## 2. 모집단과 표본

- 모집단과 표본

- Sampling : 표본추출





## 2. 모집단과 표본

### ● 모수와 통계량 표현

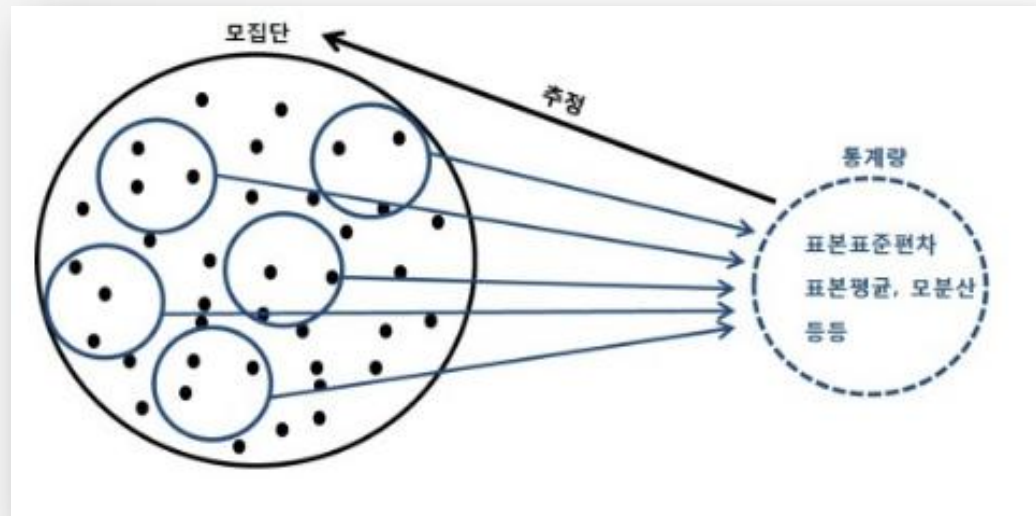
구분	모수(모집단)	통계량(표본)
의미	모집단의 특성을 나타내는 수치	표본의 특성을 나타내는 수치
표기	그리스, 로마자	영문 알파벳
평균	(모평균)	(표본의 평균)
표준편차	(모표준편차)	S (표본의 표준편차)
분산	(모분산)	$S^2$ (표본의 분산)
대상 수	N(사례수)	n(표본수)



### 3) 추정과 검정

#### ● 통계적 추정

- 모집단의 특성을 대표하는 표본을 추출하고, 이러한 표본을 이용하여 모집단의 특성을 나타내는 각종 모수(모평균, 모분산 등)를 예측하는 방법





### 3) 추정과 검정

#### ● 통계적 추정

- 모집단의 특성을 대표하는 표본을 추출하고, 이러한 표본을 이용하여 모집단의 특성을 나타내는 각종 모수(모평균, 모분산 등)를 예측하는 방법

구분	점 추정	구간 추정
방식	<ul style="list-style-type: none"><li>모집단의 특성을 <b>하나의 값</b>으로 추정하는 방식</li><li>모평균은 <b>25</b>정도로 추정</li></ul>	<ul style="list-style-type: none"><li>모집단의 특성을 <b>적절한 구간</b>을 이용하여 추정하는 방식</li><li>모평균은 <b>20~30</b> 사이로 추정</li></ul>
특징	<ul style="list-style-type: none"><li>모수와 동일할 가능성이 가장 높은 <b>하나의 값</b>을 선택하는 방법</li></ul>	<ul style="list-style-type: none"><li>모수가 속하는 <b>일정구간</b>(하한값, 상한값)으로 추정(일반적으로 많이 사용)</li></ul>



### 3) 추정과 검정

#### ● 구간추정 주요 용어

- 신뢰수준(Certainty Level) : 계산된 구간이 모수를 포함할 확률 의미 (통상 90%, 95%, 99% 등으로 표현)
- 신뢰구간(Certainty Interval) : 신뢰수준 하에서 모수를 포함하는 구간 (하한값 ~ 상한값 형식으로 표현)
- 표본오차(Sampling Error) : 모집단에서 추출한 표본이 모집단의 특성과 정확히 일치하지 않아서 발생하는 확률의 차이

예)) 대통령 후보의 지지율 여론조사에서 모 후부의 지지율이 95% 신뢰수준에서 표본오차  $\pm 3\%$  범위에서 32.4%로 조사 되었다고 가정한다면 실제 지지율은 29.4%~35.4%(-3%~+3%)사이에 나타날 수 있다는 의미이다. 여기서 95% 정도는 이 범위의 지지율을 신뢰할 수 있지만 5% 수준에서는 틀릴 수도 있는 의미이다.

➔ **신뢰수준 95%, 신뢰구간 29.4%~35.4%, 표본오차  $\pm 3\%$ ,**

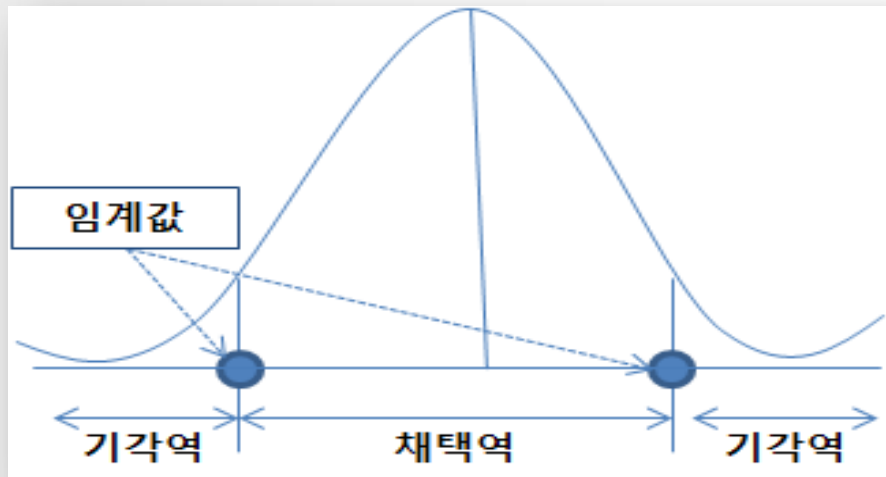




### 3) 추정과 검정

#### ● 임계값에 따른 기각역과 채택역

- 임계값(Critical value) : 귀무가설 채택 or 기각 기준점
- 채택역(Acceptance region) : 임계값 기준 채택(귀무가설) 범위
- 기각역(Critical region) : 기각 범위





### 3) 추정과 검정

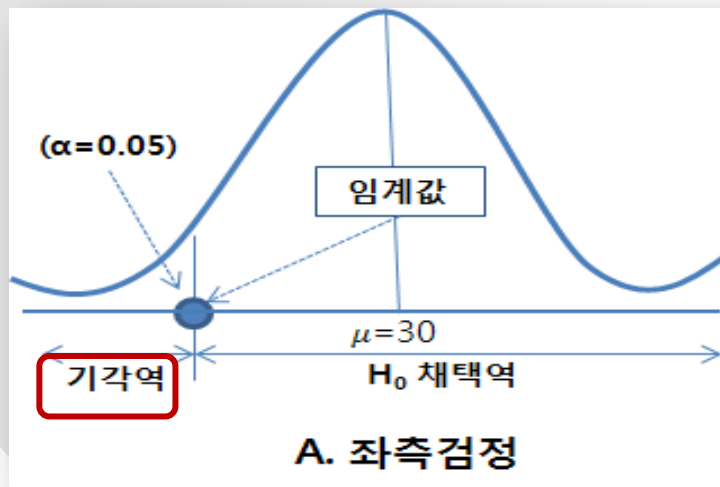
- 단측검정(1-sided test) : 방향(우열) 있는 단측가설 검정

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

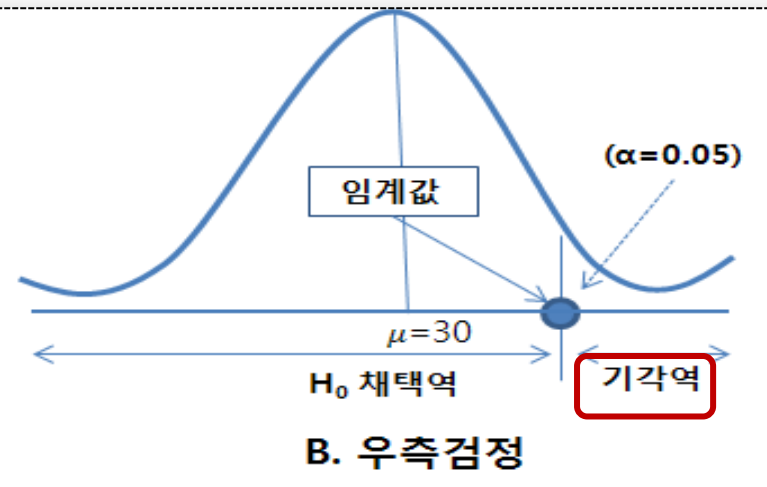
$H_1$  : 1일 생산되는 불량품의 개수는 평균 30개 이하이다. ( $\mu < 30$ ) ▶ 왼쪽 단측검정

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

연구가설이 < 또는 > 두 가지 가설 포함



왼쪽 단측검정



오른쪽 단측검정



### 3) 추정과 검정

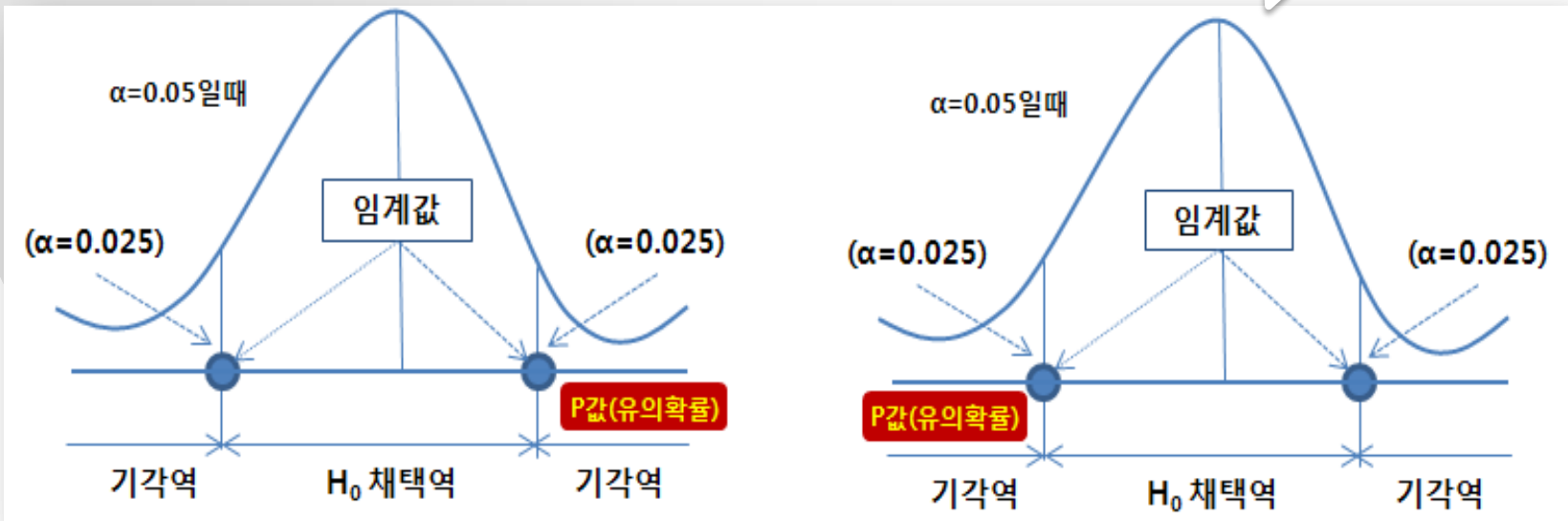
- 양측검정(2-sided test) : 방향 없는 양측가설 검정

$H_0$  : 성별에 따라 만족도에 차이가 없다.(같다)

$H_1$  : 성별에 따라 만족도에 차이가 있다.(같지 않다)

대립가설  
연구 환경에  
따라 달라짐

3가지 대립가설 : 같지 않다. 남자 > 여자, 남자 < 여자

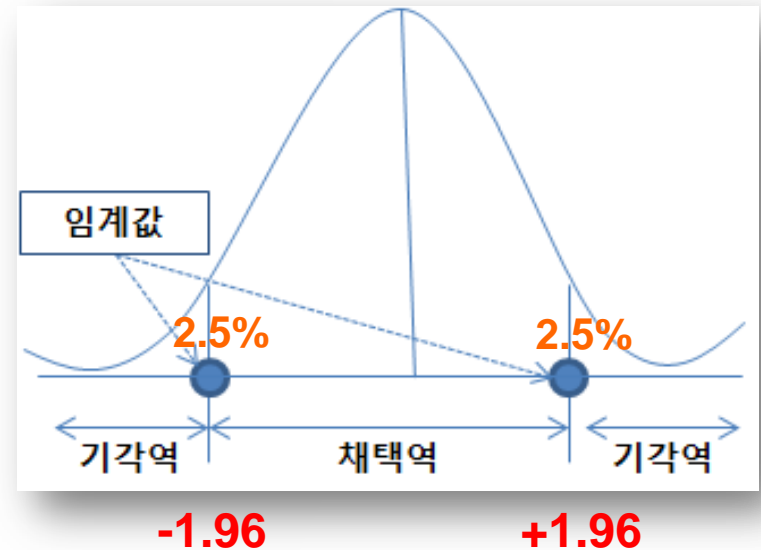




### 3) 추정과 검정

- 유의수준 vs Z값(채택역)

유의수준( $\alpha$ )/확률	정규분포 Z값(채택역)
0.5%(0.005)/99%	$\pm 2.58$ (양측검정)
2.5%(0.025)/95%	$\pm 1.96$ (양측검정)
5%(0.05)/90%	$\pm 1.64$ (양측검정)





### 3) 추정과 검정

#### ● T 분포표

Z 분포 이용 :  
모집단의  
표준편차가  
알려진 경우

T 분포 이용 :  
모집단의  
표준편차가  
알려지지 않은  
경우 표본  
표준편차 이용

자유도 $\nu$	표리확률 $q$									
	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	12.924	23.326	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.132	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.172	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.908
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	6.164
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.849	5.598
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.599	5.301
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.792	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
$\infty$	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

95% 신뢰수준 경우  
알파 = 0.025(좌우대칭)



## 4) 가설검정 오류

- 제1종 오류

- 귀무가설이 참인 경우 귀무가설 기각 오류

- 제2종 오류

- 귀무가설이 거짓인 경우 귀무가설 채택 오류

가설현황 검정 결과	귀무가설( $H_0$ ) 참인 경우	연구가설( $H_1$ ) 참인 경우
귀무가설( $H_0$ ) 채택	문제 없음	제2종 오류
연구가설( $H_1$ ) 채택	제1종 오류	문제 없음

- ❖ 가설검정에서 두 가지 오류 발생(모두 작은 경우가 바람직함)
- ❖ 제1종 오류가 발생하는 것을 가만해서 유의수준 정함  
(유의수준  $\alpha$  : 0.1, 0.05, 0.01)
- ❖ 제2종 오류를 범하지 않을 확률은  $1-\beta$  = 검정력(Power of the test)



## 5) 검정통계량

### ● 검정통계량(Test statistic)

- 가설 검정 위해 수집된 자료를 계산한 통계량
- 가설검정에서 기각역을 결정하는 기준이 되는 통계량
- 유의수준  $\alpha$ 의 값과 비교하여 귀무가설 기각/채택
- 상관분석  $r$ 값, T검정  $t$ 값, 분산분석/회귀분석  $F$ 값, 카이제곱  $\chi^2$ 값





## 5) 검정통계량

연구가설( $H_1$ ) : '학력수준에 따라 제품만족도에 차이가 있다.'를 검정하기 위해서 독립표본 T검정을 수행하였다. 이때 유의수준은  $\alpha=0.05$ 로 결정 하였다.

검정 결과 검정통계량 t값이 10.652, 유의확률 p값이 0.012가 나왔다고 가정한다면 귀무가설은 기각되는가? 채택되는가?

검정통계량  $t=10.652$ 값은 유의확률  $p=0.012$ 이다. 유의수준  $\alpha=0.05$  수준에서 귀무가설('학력수준에 따라 제품만족도에 차이가 없다.') **기각( $p < \alpha$ )**

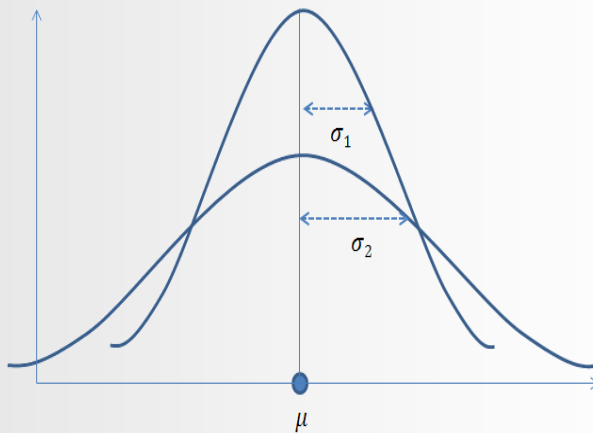
➔ 학력수준에 따라 제품만족도에 유의미한 차이가 있는 것으로 볼 수 있다.



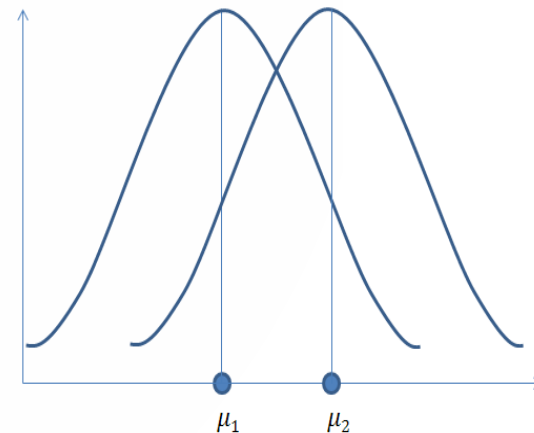
## 6) 정규분포

### ● 정규분포(Normal Distribution)

- 도수분포곡선이 평균값을 중앙으로 하여 좌우대칭인 종 모양
- K.F.가우스가 측정오차의 분포에서 중요성 강조 → 가우스분포(가우스곡선)
- 평균과 표준편차에 의해서 정규분포 모양과 위치가 결정



표준편차( $\sigma_1, \sigma_2$ )에 따른 그래프 모양



평균( $\mu_1, \mu_2$ )에 따른 그래프 모양



## 6) 정규분포

### ● 정규분포(Normal Distribution)의 특징

- 데이터의 분포가 평균을 중심으로 많은 데이터가 모여 있는 특성
- 대부분 정규분포를 이룬다고 가정하고, 통계분석 진행 → 모수 검정
- '중심극한의 정리'에 의해서 데이터의 수가 많아질수록 정규분포를 따른다.

구분	특징
변수	• 연속 변수
분포	• 평균을 중심으로 좌우대칭인 종 모양
대푯값	• 평균 = 중앙값 = 최빈값
왜도/첨도	• 왜도 = 0, 첨도 = 0(또는 3)
모양	• 표준편차( $\sigma$ )에 의해서 모양이 달라진다.
위치	• 평균( $\mu$ )에 의해서 위치가 달라진다.
넓이	• 정규분포의 전체 면적은 1이다.

※ 표준정규분포 : 평균이 0이고, 표준편차가 1인 정규분포  $N(0, 1^2)$



## 6) 정규분포

### ● 대푯값 기술통계량

- 자료 전체를 대표하는 값(분포의 중심위치를 나타내는 측정치)
- 합계(Sum), 평균(Mean)
- 중위수(Median), 최빈수(mode), 사분위수



## 6) 정규분포

- 산포도 기술통계량

- 변량이 흩어져있는 정도(평균에 모여 있으면 산포도가 작다)

- 평균( $\mu$ ) = 
$$\frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

- 분산( $\sigma^2$ ) = 
$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$

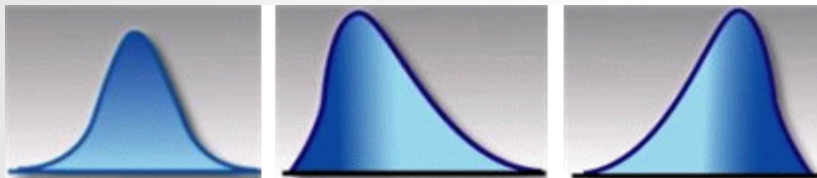
- 표준편차( $\sigma$ ) = 
$$\sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$



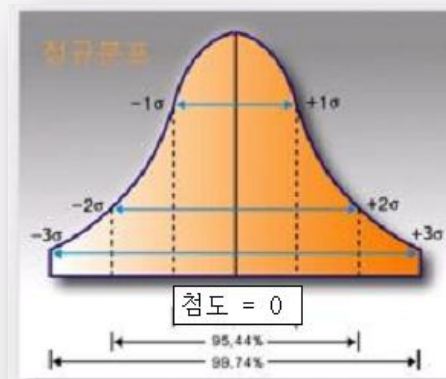
## 6) 정규분포

- 비대칭도 기술통계량

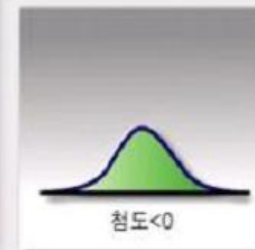
➤ 분포가 기울어진 방향과 정도



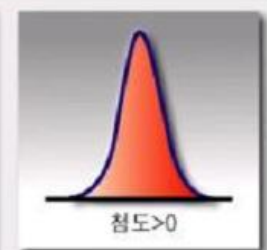
왜도=0      왜도 > 0      왜도 < 0



첨도=0



첨도 < 0



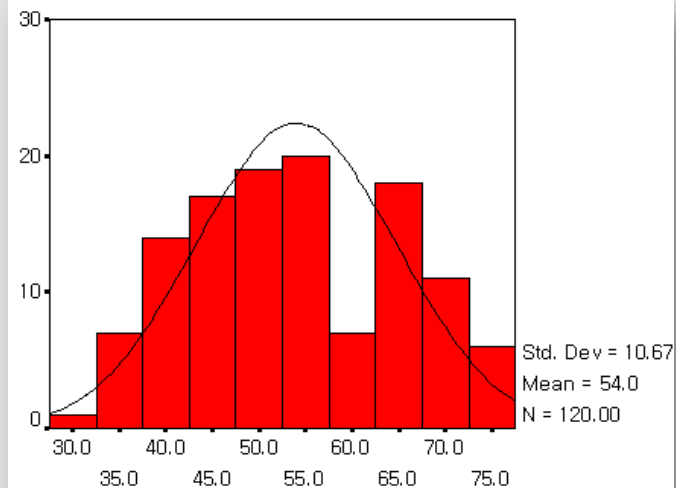
첨도 > 0



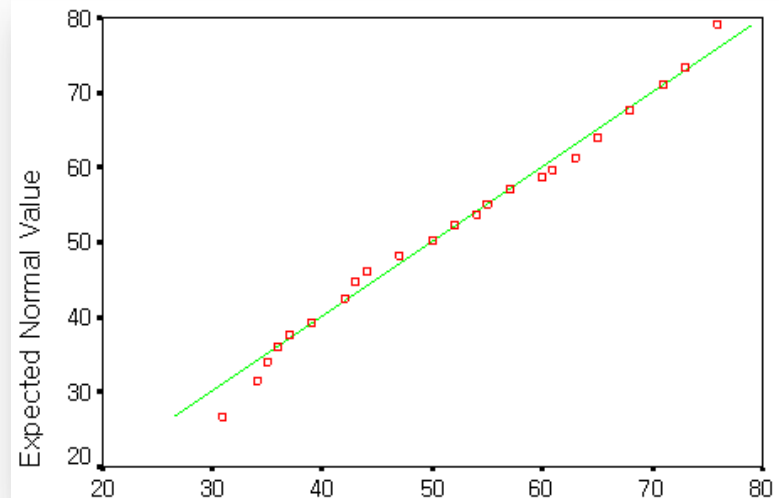
## 6) 정규분포

### ● 정규성 검정 관련 그래프

1. Graphs → Histogram



2. Graphs → Q-Q Plots







## 7) 모수 vs 비모수

- **모수(Parametric) 검정**

- 관측값이 확률분포(정규분포, 이항분포 등)를 따른 경우

- **비모수(Non-parametric) 검정**

- 관측값이 어느 특정한 확률분포를 따른다고 전제할 수 없는 경우

### 【중심극한정리】

- 케이스 30개 이상이면 정규분포를 따른다고 전제  
➔ 모수 검정 방법 실시

정규성  
검정



## 7) 모수 vs 비모수

### ● 모수 vs 비모수 검정 방법

검정 방법	모수(정규분포)	비모수(비정규분포)
t검정	독립표본 t검정	윌콕슨(Wilcoxon) 검정
	대응표본 t검정	맨-휘트니(Mann-Whitney) 검정
분산분석	일원배치분산분석	크루스칼-월리스(Kruskal-Wallis)검정
관계분석	상관분석	비모수적 상관분석



# 11. 기술통계 분석

## chap11\_DescriptiveStatistics 수업내용

- 1) 변수(변인)
- 2) 척도
- 3) 척도별 기술통계
- 4) 기술통계량 보고서



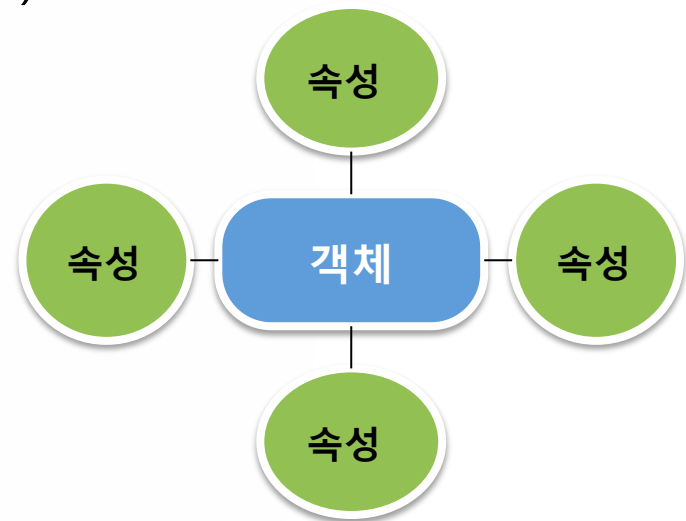
# 1) 변수(변인)

- 변수(Variable)

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

- 인구통계학적변수

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





# 1) 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.

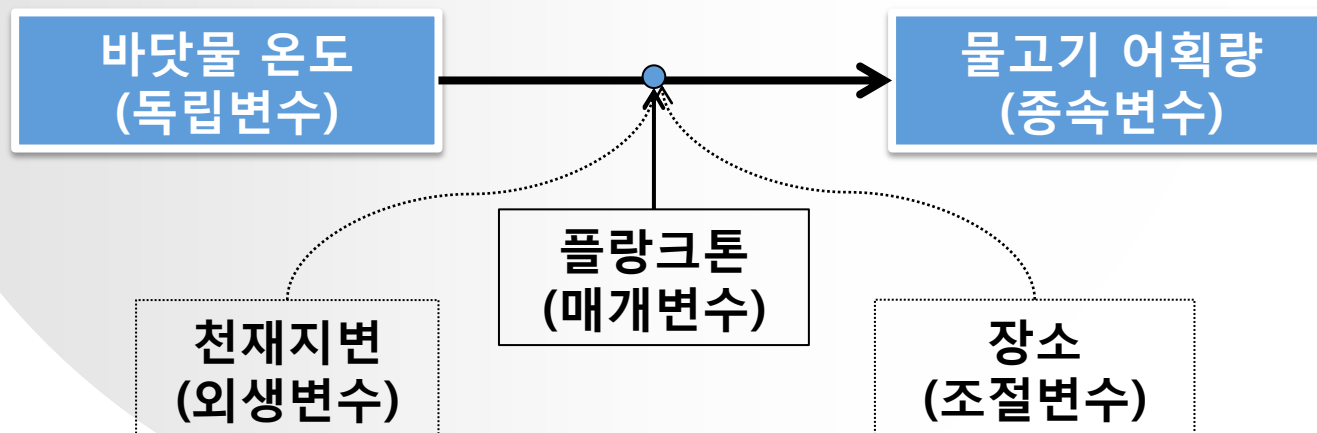
- 독립변수 : 모유 먹인 기간(예 : 3수준 척도 : ① 0 ② 1~5 ③ 5개월 이상)
- 종속변수 : 어린이와 어머니의 관계에 대한 가까운 정도
- 통제변수 : 2살 된 어린이



# 1) 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 연결시켜주는 변수
- ④ 조절변수 : 독립변수와 종속변수간 관계의 강도를 조절해주는 변수
- ⑤ 외생변수 : 독립변수와 종속변수의 관계를 잘못 이해 하게 만드는 변수





## 2) 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)	
<b>명목척도</b>	이름이나 범주를 대표하는 의미 없는 숫자 (예 : ① 남자 ② 여자)	<b>등간척도</b>	속성에 대한 각 수준 간의 간격이 동일한 경우(가감산 연산) (예: 연소득이 어디에 해당되십니까?)
<b>서열척도</b>	측정 대상 간의 높고 낮음(서열), 순서에 대한 값 부여 (예 : 좋아하는 순위를 표시하시오.)	<b>비율척도</b>	등간척도의 특성에 절대원점(0)이 존재하고, 비율계산이 가능한 경우(사칙연산) (예 : 나이가 몇 세 입니까?)



## 2) 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 등

설문지 예문) 본인의 최종학력을 표시하십시오.

① 초졸 ② 중졸 ③ 고졸 ④ 대졸 ⑤ 대학원졸





## 2) 척도

### ● 서열척도(Ordinal scale)

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

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

커피( )    녹차( )    홍차( )    우유( )



## 2) 척도

### ● 등간척도(Interval scale)

- 측정대상의 속성에 대한 각 수준 간의 간격이 동일한 척도
- 덧셈과 뺄셈 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.



## 2) 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



## 2) 척도

### ● 통계분석 방법과 변수척도 관계

분석방법	적용분야	변수척도
빈도분석	가장 기초적이고 간단한 분석방법	<b>모든 척도</b>
교차분석 (카이제곱)	변수 간의 교차표 작성	명목척도, 서열척도
요인분석	<ul style="list-style-type: none"> <li>타당성 검정</li> <li>설명력 부족한 변수 제거</li> </ul>	<b>등간척도,비율척도</b>
신뢰도분석	추출된 요인들의 동질적인 변수 구성	<b>등간척도,비율척도</b>
상관관계분석	측정변수들 간의 관계 정도를 제시	<b>피어슨 - 등간척도, 비율척도</b>
		스피어만 - 서열척도
회귀분석	인과관계 분석	<b>독립변수, 종속변수 : 등간척도/비율척도</b>
t-검정	집단 간 평균 차이 검정	독립변수 : 명목척도 종속변수 : <b>등간척도 또는 비율척도</b>
분산분석 (ANOVA)	3집단 이상의 평균 검정	독립변수 : 명목척도 종속변수 : <b>등간척도 또는 비율척도</b>



### 3) 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



### 3) 척도별 기술통계

- 척도 유형

resident	gender	age	level	cost	type	survey	pass
거주지역	성별	나이	학력수준	생활비	학교유형	만족도	합격여부
명목	명목	이율	서열	비율	명목	등간	명목
1~3	1,2	25~75	1,2,3	5.4	1,2	5점	1,2



### 3) 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)`



### 3) 척도별 기술통계

- 명목척도 변수의 기술통계량

# 명목상 의미 없는 수치로 표현된 변수 - 성별(gender)

**length(data\$gender)**

**summary(data\$gender)** # 최소, 최대, 중위수, 평균-의미없음

**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 만듦

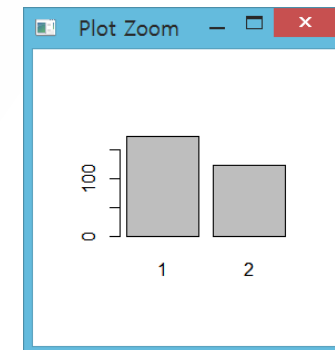
**barplot(x)** # 범주형(명목/서열척도) 시각화 -> 막대차트

**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





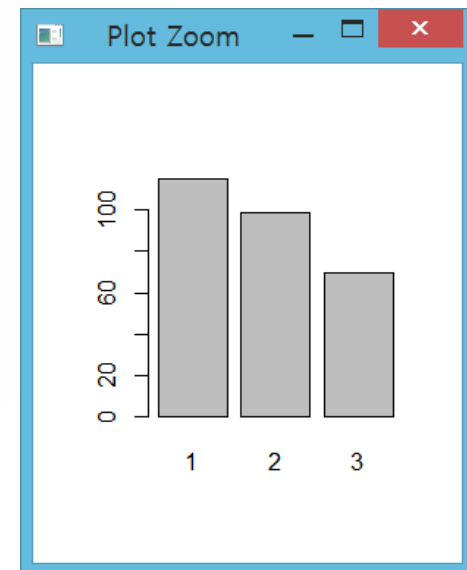


### 3) 척도별 기술통계

- 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





### 3) 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
```

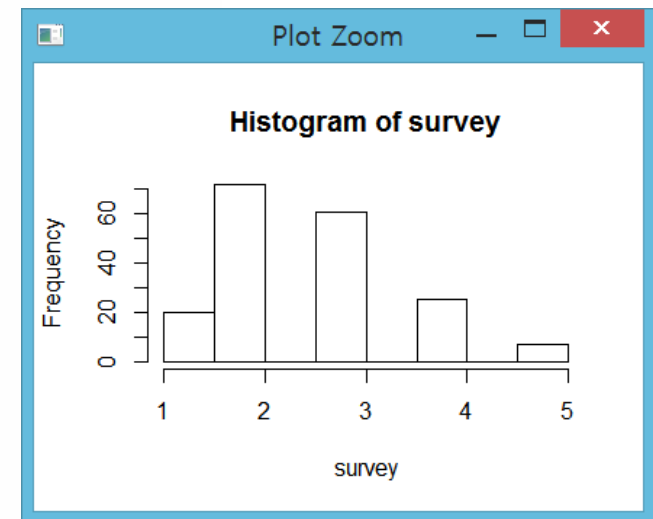
```
x1
```

```
#1 2 3 4 5
```

```
#20 72 61 25 7
```

```
hist(survey)
```

# 연속형 척도 시각화 -> 범주화 -> 히스토그램





### 3) 척도별 기술통계

- 비율척도 변수의 기술통계량

```
# 수치로 직접 입력한 변수(cost)
```

```
length(data$cost)
```

```
summary(data$cost) # 요약통계량 - 의미있음(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
```



### 3) 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



### 3) 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 뽕족한가 측정 지표
```

```
kurtosis(cost) # 2.683438
```

```
# 표준정규분포와 비교하여 첨도가 3이며 정규분포 곡선을 이루고,
```

```
# 첨도가 3보다 크면 정규분포 보다 뽕족한 형태, 3보다 작으면
```

```
# 정규분포 보다 완만한 형태이다.
```

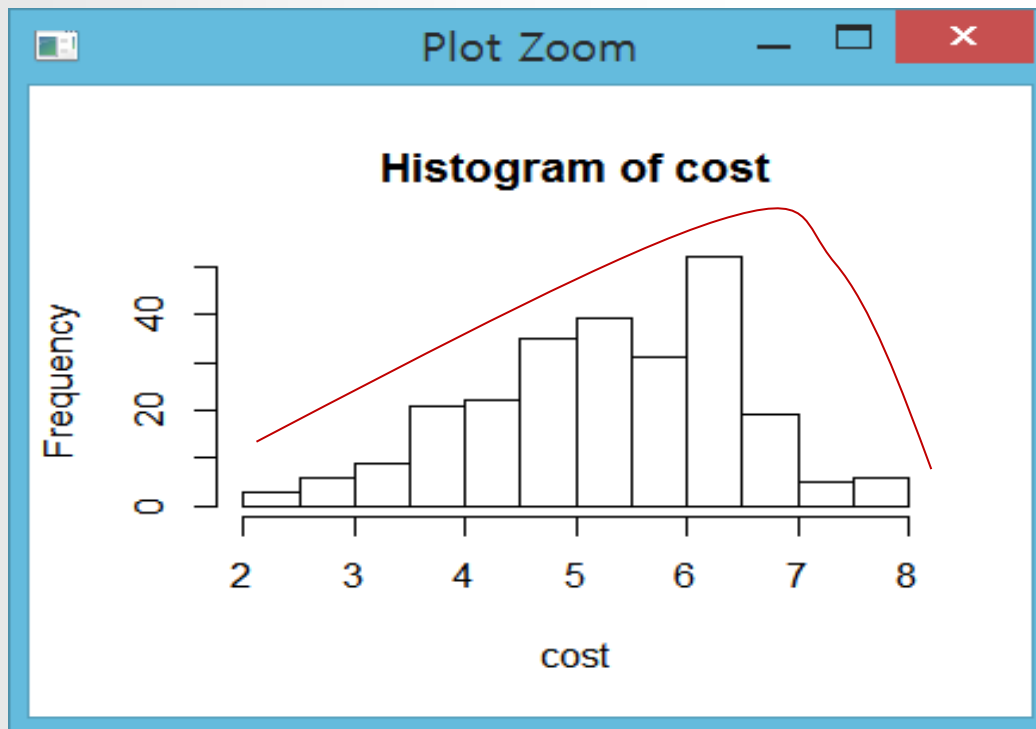
```
hist(cost) # 히스토그램으로 왜도/첨도 확인
```

```
# 왼쪽방향 비대칭 꼬리, 정규분포 첨도 보다 완만함
```



### 3) 척도별 기술통계

- 왜도/첨도에 의한 비대칭도 시각화



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



### 3) 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)`



### 3) 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
```





### 3) 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
```



### 3) 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
```



### 3) 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
```



### 3) 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
```



### 3) 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>



## 4) 기술통계량 보고서

❖ 논문에서 응답자의 인구통계적특성은 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%를 차지하였다. 또한 자녀의 대학진학여부에서 합격은 139명으로 59.15%를 차지하고, 실패는 96명으로 40.85%를 차지한 것으로 나타났다.

-----



## 4) 기술통계량 보고서

표본의 인구통계적 특성 결과

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



## 12. 교차분석과 chi-square 분석

### chap12\_CrossTableChiSquare 수업내용

#### 1) 교차표 작성/분석

- `data.frame()` 이용 교차표 작성
- `package` 이용 교차표 작성
- 교차표 분석(학력수준과 진학 여부 교차분석)

#### 2) Chi-square 가설검정

- 교차분석/ Chi-square 보고서 작성법
- ① 적합성 검정
  - ② 독립성 검정
  - ③ 동질성 검정





# 1) 교차표 작성/분석

## ● data.frame() 이용 교차표 작성

```
setwd("c:/Rwork/Part-III")
```

```
data <- read.csv("cleanDescriptive.csv", header=TRUE)
```

```
data # 확인
```

```
head(data) # 변수 확인
```

```
x <- data$level2 # 학력수준 리코딩 변수
```

```
y <- data$pass2 # 대학진학 리코딩 변수
```

```
# 학력수준(독립변수) -> 진학여부(종속변수)
```

```
result <- data.frame(Level=x, Pass=y) # 데이터 프레임 생성 - 데이터 묶음
```

```
dim(result) # 차원보기 -> 248 2
```

```
table(result) # 교차표 보기
```

```
#           Pass
# Level   실패 합격
# 고졸    40  49
# 대졸    27  55
# 대학원졸 23  31
```



# 1) 교차표 작성/분석

- package 이용 교차표 작성

# 교차표 작성을 위한 패키지 설치

```
install.packages("gmodels")
```

```
library(gmodels) # CrossTable() 함수 사용
```

# diamonds 데이터 사용을 위한 ggplot2 패키지 설치

```
install.packages("ggplot2")
```

```
library(ggplot2)
```

# diamond의 cut과 color에 대한 교차표 생성

```
CrossTable(x=diamonds$color, y=diamonds$cut, chisq = TRUE)
```





# 1) 교차표 작성/분석

## ● package 이용 교차표 작성

Total Observations in Table: 53940

제목 없음 - 메모장						
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)						
diamonds\$color	diamonds\$cut					
	Fair	Good	Very Good	Premium	Ideal	Row Total
D	163	662	1513	1603	2834	6775
	7.607	3.403	0.014	9.634	5.972	
	0.024	0.098	0.223	0.237	0.418	0.126
	0.101	0.135	0.125	0.116	0.132	
	0.003	0.012	0.028	0.030	0.053	
E	224	933	2400	2337	3903	9797
	16.009	1.973	19.258	11.245	0.032	
	0.023	0.095	0.245	0.239	0.398	0.182
	0.139	0.190	0.199	0.169	0.181	
	0.004	0.017	0.044	0.043	0.072	
F	312	909	2164	2331	3826	9542
	2.596	1.949	0.333	4.837	0.049	
	0.033	0.095	0.227	0.244	0.401	0.177
	0.194	0.185	0.179	0.169	0.178	
	0.006	0.017	0.040	0.043	0.071	
G	314	871	2299	2924	4884	11292
	1.575	23.708	20.968	0.473	30.745	
	0.028	0.077	0.204	0.259	0.433	0.209
	0.195	0.178	0.190	0.212	0.227	
	0.006	0.016	0.043	0.054	0.091	
H	303	702	1824	2360	3115	8304
	12.268	3.758	0.697	26.432	12.390	
	0.036	0.085	0.220	0.284	0.375	0.154
	0.188	0.143	0.151	0.171	0.145	
	0.006	0.013	0.034	0.044	0.058	
I	175	522	1204	1428	2093	5422
	1.071	1.688	0.090	1.257	2.479	
	0.032	0.096	0.222	0.263	0.386	0.101
	0.109	0.106	0.100	0.104	0.097	
	0.003	0.010	0.022	0.026	0.039	
J	119	307	678	808	896	2808
	14.772	10.427	3.823	11.300	45.486	
	0.042	0.109	0.241	0.288	0.319	0.052
	0.074	0.063	0.056	0.059	0.042	
	0.002	0.006	0.013	0.015	0.017	
Column Total	1610	4906	12082	13791	21551	53940
	0.030	0.091	0.224	0.256	0.400	

175  
수정



# 1) 교차표 작성/분석

- **학력수준과 대학진학여부 교차분석(Package 이용)**

# 학력수준(독립변수) : y -> 진학여부(종속변수) : x

# 학력수준이 대학 진학에 영향을 미친다.

x <- data\$level2 # 행 - 리코딩 변수 이용

y <- data\$pass2 # 열 - 리코딩 변수 이용

CrossTable(x,y) # x:학력수준, y:대학진학



# 1) 교차표 작성/분석

## ● 부모의 학력수준과 자녀의 대학진학 여부

Total Observations in Table: 225

x	y	실패	합격	Row Total
고졸		40	49	89
		0.544	0.363	0.396
		0.449	0.551	
		0.444	0.363	
		0.178	0.218	
대졸		27	55	82
		1.026	0.684	0.364
		0.329	0.671	
		0.300	0.407	
		0.120	0.244	
대학원졸		23	31	54
		0.091	0.060	0.240
		0.426	0.574	
		0.256	0.230	
		0.102	0.138	
Column Total		90	135	225
		0.400	0.600	

▪ 기대치 비율 예 (1행1열)

▪ 기대치 :  $89(\text{관측치}) \times 0.4(\text{행비율}) = 35.6$

▪ 기대치 비율 :  $(40-35.6)^2/35.6 = \mathbf{0.5438}$

관측치

기대치비율( $\chi^2$ ) =  $(\text{관측치} - \text{기대치})^2 / \text{기대치}$

행비율

열비율

셀비율

관측치

$(\text{관측치} - \text{기대치})^2 / \text{기대치}$

행비율

열비율

셀비율

관측치

$(\text{관측치} - \text{기대치})^2 / \text{기대치}$

행비율

열비율

셀비율

전체 관측치

전체 열비율



# 1) 교차표 작성/분석

❖ 논문에서 교차분석에 대한 해설 예

<교차분석 해설>-----

부모의 학력수준에 따른 자녀의 대학진학여부를 설문조사한 결과 학력수준에 상관없이 대학진학 합격률이 평균 59.6%로 학력수준별로 유사한 결과가 나타났다. 전체 응답자 228명을 대상으로 고졸 39.9% (89명) 중 55.1%가 진학에 성공하였고, 대졸 36.4%(82명) 중 68.4%가 성공했으며, 대학원졸은 24%(54명) 중 57.4%가 대학진학에 성공하였다. 특히 대졸 부모의 대학진학 합격율이 평균보다 조금 높고, 고졸 부모의 대학진학 합격율이 평균보다 조금 낮은 것으로 분석된다.

-----



## 2) Chi-square 검정

### ● Chi-square 검정

- 범주(Category)별로 관측 빈도와 기대빈도가 차이가 있는지 검정
- 카이제곱 분포에 기초한 통계적 방법(카이제곱 분포표 이용)
- $\chi^2 = \sum (\text{관측값} - \text{기댓값})^2 / \text{기댓값}$
- 분석을 위해서 교차분할표 작성
- 교차분석은 검정통계량으로 카이제곱 사용(=카이제곱 검정)
- 검증 유형 분류 : 일원카이제곱검정, 이원카이제곱검정



## 2) Chi-square 검정

1. 일원카이제곱 : 교차분할표 이용 안함(한 개 변인)
  - 적합성 검정 : 실제 표본이 내가 생각하는 분포와 같은가? 다른가?  
예) 관찰도수가 기대도수와 일치하는지를 검정
2. 이원카이제곱 : 교차분할표 이용
  - 1) 독립성 검정 : **두 변인**은 서로 관련성이 있는가 없는가?
    - 한 모집단으로부터 하나의 표본이 추출된 경우
    - 예) 흡연량과 음주량 사이에 관련성이 있는가?
    - 귀무가설 : 흡연과 음주량은 관련성이 없다.(독립적이다.)
  - 2) 동일성 검정 : **두 집단**의 분포가 동일한가? 다른 분포인가?
    - 두 개 이상의 범주형 자료가 동일한 분포를 갖는 모집단에서 추출된 것인지 검정하는 방법
    - 두 개 이상의 모집단에서 각 표본이 추출된 경우
    - 귀무가설 : 집단 간의 비율이 동일하다.





## 2) Chi-square 검정

### ● Chi-square 검정 절차

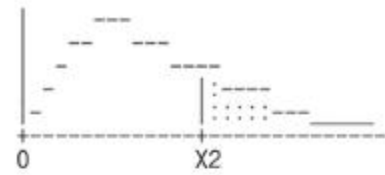
1. 가설을 설정한다.
2. 유의수준을 결정한다.
3. 기각값(카이제곱 분포표 참조)을 결정한다.
  - 자유도(df)와 유의수준으로 기각값 결정
4. 관찰도수에 대한 기대도수를 구한다.
5. 검정통계량  $\chi^2$ 의 값을 구한다.
6. 귀무가설의 채택 또는 기각 여부를 판정한다.
7. 카이제곱 검정 결과를 설명한다.



## 2) Chi-square 검정

### ● 카이제곱 분포표

CHI-SQUARE TABLE: VALUES OF CHI-SQUARE (ALPHA) OF THE CHI-SQUARE DISTRIBUTION



자유도

유의수준

DF	X2( .995)	X2( .99)	X2( .975)	X2( .95)	X2( .05)	X2( .025)	X2( .01)	X2( .005)
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928

자유도 =  $n-1$   
( $n$ 은 표본수)



## 2) Chi-square 검정

### 1. 일원카이제곱 검정

#### (1) 적합성 검정 - `chisq.test()` 이용

귀무가설 : 기대치와 관찰치는 차이가 없다.

예) 도박사의 주사위는 게임에 적합하다.

대립가설 : 기대치와 관찰치는 차이가 있다.

예 도박사의 주사위는 게임에 적합하지 않다.

# 주사위의 관찰치가 기대치와 차이가 있는가? 또는 없는가?

# 60회 주사위를 던져서 나온 관측도수/기대도수

# 관측도수 : 4, 6, 17, 16, 8, 9

# 기대도수 : 10, 10, 10, 10, 10, 10

`chisq.test(c(4,6,17,16,8,9))` # p-value = 0.01439

# 해설 : 도박사의 주사위는 게임에 적합하지 않다.



## 2) Chi-square 검정

- p값 해석 방법

<해설> p값이 0.05미만이기 때문에 유의미한 수준에서 귀무가설을 기각할 수 있다.  
따라서 '도박사의 주사위는 게임에 적합하지 않다.'라는 대립가설을 채택한다.  
(귀무가설 기각, 대립가설 채택)

- 유의수준과 유의확률

- # 유의수준(Confidence level) : 0.05(100개 중 5개( $100 \times 0.05$ ) 허용 기준치(허용 오차)
- # 유의확률 : p-value 귀무가설이 나올 수 있는 확률
- # p-value < 0.05 경우 : 유의확률은 유의수준 보다 적다.(귀무가설 기각)

- 검정통계량 해석 방법

- # 검정통계량 : X-squared = 14.2, df = 5
- # 자유도(df) : 관측치가 n 인 경우  $df = n - 1$
- # 자유도(degree of freedom)란 검정을 위해서 n개의 표본(관측치)를 선정할 경우
- # n번째 표본은 나머지 표본이 정해지면 자동으로 결정되는 변인의 수를 의미
- # 자유도(df) 5인 경우, X-squared 기각값(역) :  $\chi^2 \geq 11.071$  (chi-square 분포표 참고)
- #  $\chi^2$  값이 11.071 이상이면 귀무가설을 기각할 있다는 의미



## 2) Chi-square 검정

### (2) 선호도 분석

귀무가설 : 기대치와 관찰치는 차이가 없다.

예) 맥주의 선호도에 차이가 없다.

대립가설 : 기대치와 관찰치는 차이가 있다.

예) 맥주의 선호도에 차이가 있다.

```
data <- textConnection(
```

```
"맥주종류 관측도수
```

```
1 12
```

```
2 30
```

```
3 15
```

```
4 7
```

```
5 16")
```

```
x <- read.table(data, header=T)
```

```
chisq.test(x$관측도수) # X-squared = 18.375, p-value = 0.001042
```

```
# 해설 : 맥주의 선호도에 차이가 있다.
```



## 2) Chi-square 검정

- 선호도 분석 결과

- 검정통계량 :

$$\chi^2 = 18.375, df = 4$$

- p-value 해석 :

p값이 0.05미만이기 때문에 유의미한 수준에서 귀무가설을 기각할 수 있다. 따라서 '맥주의 선호도에 차이가 있다.'라는 대립가설을 채택할 수 있다. (귀무가설 기각, 대립가설 채택)



## 2) Chi-square 검정

### 2. 이원카이제곱 검정

#### 1) 독립성 검정(관련성 검정) - 교차테이블 이용

귀무가설 : 부모의 학력수준과 자녀의 대학진학 여부와 관련성이 없다.

- 두 변인은 독립적이다.

대립가설 : 부모의 학력수준과 자녀의 대학진학 여부와 관련성이 있다.

- 두 변인은 독립적이지 않다.

```
CrossTable(x, y, chisq = TRUE) # p = 0.2507057
```



## 2) Chi-square 검정

### ● 독립성 검정(관련성 검정) 결과

x	y	실패	합격	Row Total
고졸		40	49	89
		0.544	0.363	
		0.449	0.551	0.396
		0.444	0.363	
		0.178	0.218	
대졸		27	55	82
		1.026	0.684	
		0.329	0.671	0.364
		0.300	0.407	
		0.120	0.244	
대학원졸		23	31	54
		0.091	0.060	
		0.426	0.574	0.240
		0.256	0.230	
		0.102	0.138	
Column Total		90	135	225
		0.400	0.600	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi<sup>2</sup> = 2.766951      d.f. = 2      p = 0.2507057 |

### <검정 결과 해설>

- ✓  $\text{Chi}^2 = \sum [(\text{관측값} - \text{기댓값})^2 / \text{기댓값}]$
- ✓  $\text{d.f.} = (\text{행수}-1) * (\text{열}-1) = (3-1) * (2-1) = 2$   
-> 두 값만 구하면 나머지는 저절로 구해진다.
- ✓  $p = \text{유의수준} : 0.05$ 이하이면 귀무가설 기각

- # 자유도에 따른  $\text{Chi}^2$  분포도  
-> 자유도가 클 수록 정규분포에 가까워진다.
- # 유의수준 0.05에서,  
-> 자유도 : 2인 경우, 기각역 :  $\chi^2 \geq 5.99$ ,  
-> 자유도 : 6인 경우, 기각역 :  $\chi^2 \geq 12.59$
- # 자유도가 2인 경우  $\chi^2$  값이 5.99이상이면  
귀무가설 기각(카이제곱 분포표 참조)
- # 해설 :  $\text{Chi}^2$  값이 5.99 이하이고, 유의수준이 0.05 이상으로 분석되어 귀무가설을 기각할 수 없다. 따라서 부모의 학력수준과 자녀의 대학 진학 변인 간의 관련성은 없는 것으로 분석된다.





## 2) Chi-square 검정

❖ 논문에서 교차분석표와 Chi-square 검정에 대한 해설 예

<교차분석표와 카이제곱 검정결과 해설>-----

'부모의 생활수준과 자녀의 대학진학 여부와 관련성이 있다.'를 분석하기 위해서 자녀를 둔 A회사 225명의 부모를 표본으로 추출한 후 설문조사하여 교차분석과 카이제곱 검정을 실시하였다.

분석결과를 살펴보면 부모의 생활수준과 자녀의 대학진학 여부의 관련성은 유의미한 수준에서 차이가 없는 것으로 나타났다. ( $X^2=2.767$ ,  $p>0.05$ ) 따라서 귀무가설을 기각할 수 없다. 다음 <표>에서 부모의 생활 수준과 자녀의 대학 진학 여부에 대한 교차표와 카이제곱 검정결과를 제시하고 있다.

-----



## 2) Chi-square 검정

### <논문에서 카이제곱 검정 결과 제시방법>

카이제곱 검정결과를 논문에서 제시할 경우 교차표와 카이제곱 검정통계량 함께 제시

학력수준		실패	진학	X-squared	유의확률(p)
고졸	관찰빈도	40	49	2.766951	0.2507057
	기대빈도	36	54		
대졸	관찰빈도	27	55		
	기대빈도	33	49		
대학원졸	관찰빈도	23	31		
	기대빈도	21	32		



## 2) Chi-square 검정

### <실습> 교육수준과 흡연율 간의 관련성 분석

#### 1. 파일 가져오기

```
setwd("c:/Rwork/Part-III")
```

```
smoke <- read.csv("smoke.csv", header=TRUE)
```

```
# 변수 보기
```

```
head(smoke) # education, smoking 변수
```

```
names(smoke)
```

```
[1] "education" "smoking"
```

#### ● 변수 모델링

객체를 대상으로 분석할 속성(변수)을 선택하여 속성 간의 관계 설정 과정

예) smoke 객체에서 education, smoking 속성을 분석대상으로 하여 교육수준이 흡연율과 관련성이 있는가를 education -> smoking 형태로 기술한다. education은 영향을 미치는 변수로 독립변수라 하며, 영향을 받는 smoking은 종속변수라고 한다.



## 2) Chi-square 검정

### 2. 코딩 변경 - 변수 리코딩 <- 가독성 제공

# education(독립변수) : 1:대졸, 2:고졸, 3:중졸

# smoke(종속변수): 1:과다흡연, 2:보통흡연, 3:비흡연

```
table(smoke$education, smoke$smoking)
```

```
smoke$education2[smoke$education==1] <- "대졸"
```

```
smoke$education2[smoke$education==2] <- "고졸"
```

```
smoke$education2[smoke$education==3] <- "중졸"
```

```
smoke$smoking2[smoke$smoking==1] <- "과다흡연"
```

```
smoke$smoking2[smoke$smoking==2] <- "보통흡연"
```

```
smoke$smoking2[smoke$smoking==3] <- "비흡연"
```

```
smoke # 가독성을 위한 변수값 변경 결과
```



## 2) Chi-square 검정

### 3. 교차표 작성

```
table(smoke$education2, smoke$smoking2)
```

과대흡연   보통흡연   비흡연

고졸	22	21	9
대졸	51	92	68
중졸	43	28	21



## 2) Chi-square 검정

### 4. 독립성 검정

**library(gmodels) # CrossTable() 함수 사용**

**CrossTable(smoke\$education2, smoke\$smoking2, chisq = TRUE)**

Pearson's Chi-squared test

-----  
Chi^2 = 18.91092      d.f. = 4      p = 0.0008182573



## 2) Chi-square 검정

### 2) 동질성 검정 - 교차테이블 이용

귀무가설 : 집단 간의 비율이 동일하다.

예) 교육방법에 따른 만족도에 차이가 없다.

대립가설 : 집단 간의 비율이 동일하지 않다.

예) 교육방법에 따른 만족도에 차이가 있다.



## 2) Chi-square 검정

### 1. 파일 가져오기

```
setwd("c:/Rwork/Part-III")
```

```
data <- read.csv("homogeneity.csv", header=TRUE)
```

```
head(data) # 변수 보기
```

```
data <- subset(data, !is.na(survey), c(method, survey))
```





## 2) Chi-square 검정

### 2. 변수리코딩 - 코딩 변경

# method: 1:방법1, 2:방법2, 3:방법3

# survey: 1:매우만족, 2:만족, 3:보통, 4: 불만족, 5: 매우불만족

# 교육방법2 필드 추가

```
data$method2[data$method==1] <- "방법1"
```

```
data$method2[data$method==2] <- "방법2"
```

```
data$method2[data$method==3] <- "방법3"
```

# 만족도2 필드 추가

```
data$survey2[data$survey==1] <- "매우만족"
```

```
data$survey2[data$survey==2] <- "만족"
```

```
data$survey2[data$survey==3] <- "보통"
```

```
data$survey2[data$survey==4] <- "불만족"
```

```
data$survey2[data$survey==5] <- "매우불만족"
```



## 2) Chi-square 검정

### 3. 교차분할표 작성

`table(data$method2, data$survey2) # 교차표 생성 -> table(행,열)`

만족 매우만족 매우불만족 보통 불만족

방법1    8        5        6    15     16    -> 50

방법2   14        8        6    11     11    -> 50

방법3    7        8        9    11     15    -> 50

# 주의 : 반드시 각 집단별 길이(50)가 같아야 한다.



## 2) Chi-square 검정

### 4. 동질성 검정 - 모수 특성치에 대한 추론검정

**chisq.test(data\$method2, data\$survey2)**

Pearson's Chi-squared test

data: data\$method2 and data\$survey2

X-squared = 6.5447, df = 8, p-value = 0.5865

#### <해설>

유의수준 0.05에서  $\chi^2$ 값이 6.545, 자유도 8, 그리고 유의확률 0.586을 보이고 있다. 즉 6.545 이상의 카이제곱값이 얻어질 확률이 0.586라는 것을 보여주고 있다.

이 값은 유의수준 0.05보다 크기 때문에 귀무가설을 기각할 수 없다. 따라서 '교육방법에 따른 만족도에 차이가 없다.'라고 할 수 있다.



# 13. 집단 간 차이 분석

## chap13\_Ttest\_Anova 수업내용

- 1) 단일 집단 분석
- 2) 두 집단 분석
- 3) 세 집단 분석(분산 분석)



# 단일집단 비율검정

#####

## # 추론통계학 분석 - 1-1. 단일집단 비율검정

#####

# 방법 : 1개 집단의 비율과 기존 집단과의 비율 차이 분석

# 작업절차

# 1. 실습데이터 가져오기

# 2. 빈도수와 비율계산

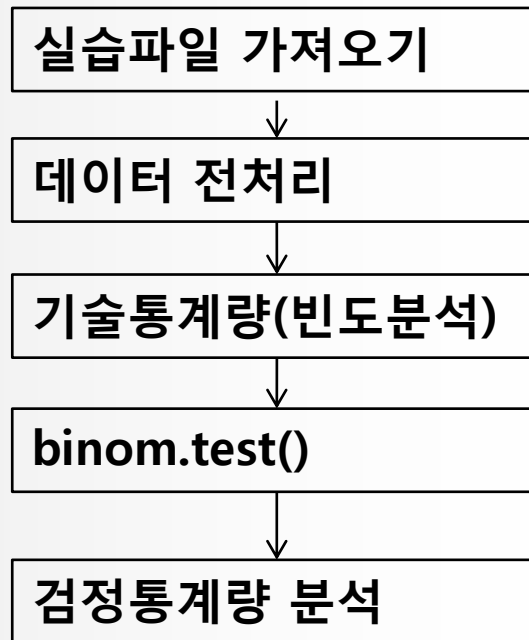
# 3. binom.test() 이용

#####



# 단일집단 비율검정

- 분석절차





# 단일집단 비율검정

## <연구가설>

- 연구가설( $H_1$ ) : 기존 2014년도 고객 불만율과 2015년도 CS교육 후 불만율에 차이가 있다.
- 귀무가설( $H_0$ ) : 기존 2014년도 고객 불만율과 2015년도 CS교육 후 불만율에 차이가 없다.

## <연구환경>

2014년도 114 전화번호 안내고객을 대상으로 불만을 갖는 고객은 20%였다. 이를 개선하기 위해서 2015년도 CS교육을 실시한 후 150명 고객을 대상으로 조사한 결과 14명이 불만을 갖고 있었다. 기존 20% 보다 불만율이 낮아졌다고 할 수 있는가?

# 대상 파일 : c:/Rwork/Part-III/one\_sample.csv

# 해당 변수 : survey(만족도)

# 변수 척도 : 명목척도(y/n)

# 가정 : 기존 불만율과 CS교육 후 불만율 분석



# 단일집단 비율검정

## 1. 실습데이터 가져오기

```
getwd()
```

```
setwd("c:/Rwork/Part-III")
```

```
data <- read.csv("one_sample.csv", header=TRUE)
```

```
head(data)
```

```
x <- data$survey # 만족도 변수
```





# 단일집단 비율검정

## 2. 빈도수와 비율 계산

**summary(x) # 결측치 없음**

**length(x) # 150개**

**table(x)**

**#x**

**# 0 1**

**# 14 136 -> 0:불만족(14), 1: 만족(136)**

**#table(x, useNA="ifany") # 시리얼 데이터와 NA 개수 출력 시**

**install.packages("prettyR")**

**library(prettyR) # freq() 함수 사용**

**freq(x)**

**# Frequencies for x**

**# 1 0 NA**

**# 136 14 0 <- 빈도수**

**#% 90.7 9.3 0 <- 비율 제공**



# 단일집단 비율검정

## 3. 가설검정 : `binom.test()` 함수 : 명목척도( $y/n$ ) 대상

### # 이항분포 개념

- # 1. 정규분포와 마찬가지로 모집단이 가지는 이상적인 분포형
- # 2. 정규분포가 연속변량, 이항분포는 이산변량
- # 3. 그래프는 좌우대칭인 종 모양 곡선

### # `binom.test()` 함수 이용 가설검정

### `help(binom.test)` # 함수 형식

```
#binom.test(x, n, p = 0.5, alternative = c("two.sided", "less", "greater"),  
#          conf.level = 0.95)
```

```
# # 형식) binom.test(만족수, 불만족수, p = 확률)
```



# 단일집단 비율검정

## 1) 만족율 기준 검정

# 양측검정

**binom.test(c(136,14), p=0.8)** # 기존 80% 만족율 기준 검증 실시

**binom.test(c(136,14), p=0.8, alternative="two.sided", conf.level=0.95)**

# alternative="two.sided" : 양측검정-> p-value = 0.0006735

# 해설 : 기존 만족율(80%)과 차이가 있다. -> 연구가설 채택

# 단측검정

**binom.test(c(136,14), p=0.8, alternative="greater", conf.level=0.95)**

# alternative="greater" : 단측검정-> 방향성 # p-value = 0.0003179

# 해설 : CS교육을 통해서 기존 만족율(80%) 이상의 효과를 얻을 수 있다고

# 볼 수 있다. 따라서 기존 20% 보다 불만율이 낮아졌다고 할 수 있다.



# 단일집단 비율검정

## 2) 불만족율 기준 검정

# 양측검정

```
binom.test(c(14,136), p=0.2) # 기존 20% 불만족율 기준 검증 실시  
binom.test(c(14,136), p=0.2, alternative="two.sided", conf.level=0.95)
```

# alternative="two.sided" : 양측검정 -> p-value = 0.0006735

# 해설 : 기존 불만족율(20%)과 차이가 있다. -> 연구가설 채택

# 단측검정

```
binom.test(c(14,136), p=0.2, alternative="greater", conf.level=0.95)
```

# alternative="greater" : 단측검정 -> 방향성 # p-value = 0.9999

# 불만족율 20% 보다 크지 않다.

```
binom.test(c(14,136), p=0.2, alternative="less", conf.level=0.95)
```

# p-value = 0.0003179 -> 불만족율 20% 보다 적다.



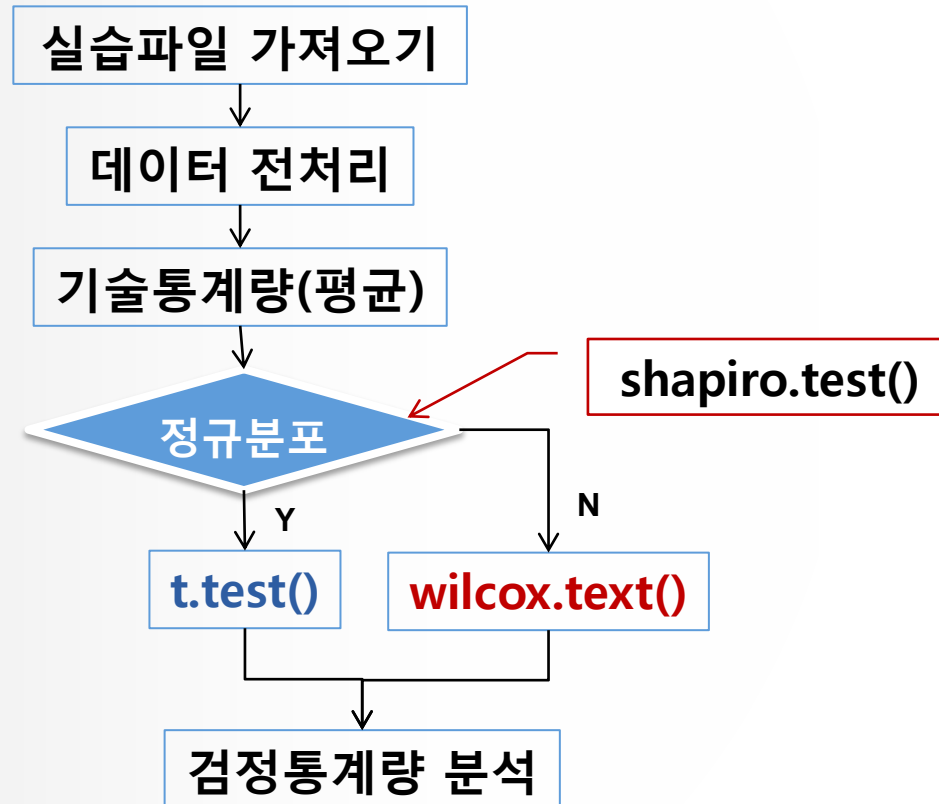
# 단일집단 평균검정

```
#####  
# 추론통계학 분석 - 1-2. 단일집단 평균 검정(단일표본 T검정)  
#####  
# 방법 : 1개 집단의 평균과 어떤 특정한 값과 차이가 있는지 검증  
# 작업절차  
# 1. 실습파일 가져오기  
# 2. 데이터 분포 및 결측치 제거(데이터 정제)  
# 3. 정규분포 검정 : 모집단의 특성 반영 유무  
# 4. 가설검정(모수/비모수) -> t.test()/wilcox.test()  
#####
```



# 단일집단 평균검정

- 분석절차





# 단일집단 평균검정

## <연구가설>

- 연구가설( $H_1$ ) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 있다.
- 귀무가설( $H_0$ ) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 없다.

## <연구환경>

국내에서 생산된 노트북 평균 사용 시간이 5.2시간으로 파악된 상황에서 A회사에서 생산된 노트북 평균 사용시간과 차이가 있는지를 검정하기 위해서 A회사 노트북150대를 랜덤으로 선정하여 검정을 실시한다.

-----  
# 대상 파일 : c:/Rwork/Part-III/one\_sample.csv

# 해당 변수 : time

# 변수 척도 : 비율척도(직접 입력한 수치 데이터)

# 가정 : 기존 노트북 평균 사용시간 vs A회사 노트북 평균 사용시간

# 검정 : 노트북 평균 사용시간 수집 -> 평균 -> 정규성 검정 -> T검정



# 단일집단 평균검정

## 1. 실습파일 가져오기

```
setwd("c:/Rwork/Part-III")
```

```
data <- read.csv("one_sample.csv", header=TRUE)
```

```
head(data)
```

```
x <- data$time # 노트북 사용 시간
```

```
head(x)
```





# 단일집단 평균검정

## 2. 데이터 분포 /결측치 제거

```
summary(x) # NA-41개
```

```
mean(x) # error
```

```
mean(x, na.rm=T) # NA 제외 평균(방법1)
```

```
# 데이터 정제 -> 5.556881
```

```
x1 <- na.omit(x) # NA 제외 평균(방법2)
```

```
X1
```

```
# 평균(mean) 특징
```

```
# 평균 모양 : 양측에 대한 균형
```

```
# 대상 : 수치 데이터 -> 비율(ratio)
```

```
# 적용 : 평균 차이 검정으로 의사결정
```

```
# 평균 검정 : 평균에 의미가 있는가 검정, 평균을 중심으로 종 모양 형성
```

```
# 왜도 : 한쪽으로 치우쳐진 정도
```



# 단일집단 평균검정

## 3. 정규분포 검정

# 정규분포(바른 분포) : 평균에 대한 검정

# 정규분포 검정 귀무가설 : 정규분포와 차이가 없다.

# shapiro학자가 만든 함수 이용 : shapiro.test()

**shapiro.test(x1) # x1 데이터에 대한 정규분포를 검정하는 함수**

**# W = 0.9914, p-value = 0.7242 <- 정규분포**

# 검정결과 분석 : 0.05보다 작으면 정규분포가 아닌 것으로 판단

# 명목적도 -> 보기 항목으로 정규분포가 그려지기 때문에 의미 없음

# 비율척도, 수치 기반 척도(평균에 의미 있는 척도) -> 정규분포 검정 필요

**# 정규분포(모수검정) -> t.test()**

**# 비정규분포(비모수검정) -> wilcox.test()**

**hist(x1) # 정규분포 형태**



# 단일집단 평균검정

## 4. 가설검정 - 모수/비모수

**# t.test()**

# - 모집단의 평균값을 검정하는 함수

# - 예) 기존평균사용시간 5.2시간 기준으로 검정(같다 vs 차이)

**help(t.test)**

# t -> student에서 t

### 1) 양측검정

**t.test(x1, mu=5.2)** # mu(그리스 로마 - 평균) : 기존 5.2시간 기준 검정

# x1 : 표본집단 평균, mu=5.2, 모집단의 평균값

# 정제 데이터와 5.2시간 비교

**t.test(x1, mu=5.2, alter="two.side", conf.level=0.95)**

# p-value = 0.0001417

# 해설 : 평균 사용시간 5.2시간과 차이가 있다.(귀무가설 기각)



# 단일집단 평균검정

## ● 점추정 vs 구간추정

#alternative hypothesis: true mean is not equal to 5.2

#95 percent confidence interval:

# 5.377613 5.736148 -> 구간추정(95% 신뢰구간 추정)

#sample estimates:

# mean of x

# 5.556881 -> 점추정 : mean값과 직접비교하여 추정

# 점추정(point) vs 구간추정(interval estimation)

# 점추정 : 모수를 하나의 값으로 추정(평균이나 중위수 사용)

# 구간추정 : 모수가 포함될 것이라고 제시하는 구간추정(신뢰구간)



# 단일집단 평균검정

## 2) 단측검정

```
t.test(x1, mu=5.2, alter="greater", conf.level=0.95)
```

```
# p-value = 7.083e-05 = 0.00007083
```

```
# 해설 : A회사 노트북의 평균 사용시간은 5.2시간 보다 더 길다.
```

```
# 검정 결과를 변수에 저장하여 특정 변수 확인하기
```

```
result <- t.test(x1, mu=5.2, alter="greater", conf.level=0.95)
```

```
names(result)
```

```
str(result)
```

```
result$p.value # 7.083346e-05 -> 세밀한 정보 제공
```



# 단일집단 평균검정

## 【단일표본 t-검정 결과 정리 및 기술】

1) 가설 설정	연구가설(H1) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 있다.
	귀무가설(H0) : 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이가 없다.
2) 연구환경	국내에서 생산된 노트북 평균 사용 시간이 5.2시간으로 파악된 상황에서 A회사에서 생산된 노트북 평균 사용시간과 차이가 있는지를 검정하기 위해서 A회사 노트북 150대를 랜덤으로 선정하여 검정을 실시한다.
3) 유의수준	$\alpha = 0.05$
4) 분석방법	단일표본 T검정
5) 검정통계량	$t = 3.9461, df = 108$
6) 유의확률	$P = 0.0001417$
7) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 국내에서 생산된 노트북과 A회사에서 생산된 노트북의 평균 사용 시간에 차이를 보인다고 할 수 있다. 즉 국내에서 생산된 노트북의 평균 사용 시간은 5.2이며, A회사에서 생산된 노트북의 평균 사용 시간은 5.56으로 국내 평균 사용 시간 보다 더 길다고 할 수 있다.



# 두 집단 비율검정

#####

# 추론통계학 분석 - 2-1. 두 집단 비율 검정

#####

# 방법 : 두 집단 간 비율 차이에 관한 분석

# 작업절차

# 1. 실습파일 가져오기

# 2. 두 집단 subset 작성(데이터 정제,전처리)

# -> 데이터 정제, 전처리

# -> 기술통계량 - 빈도수

# -> 두 변수(집단)에 대한 교차분석

# 3. 두 집단 비율차이 검정

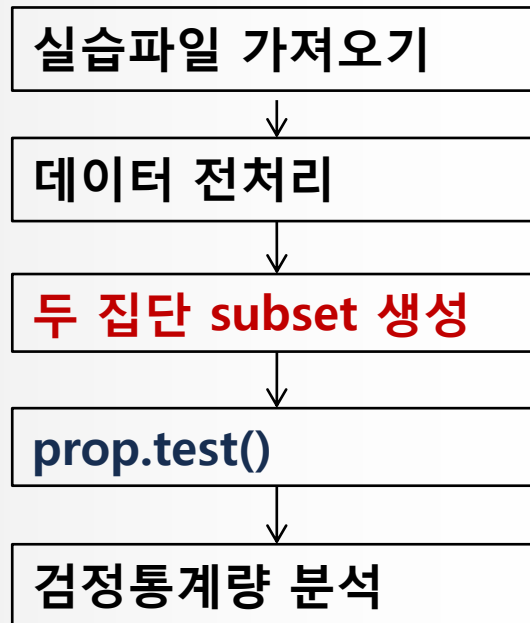
# -> prop.test()

#####



# 두 집단 비율검정

- 분석절차







# 두 집단 비율검정

## <연구가설>

- 연구가설( $H_1$ ) : 두 가지 교육방법에 따라 교육생의 만족율에 차이가 있다.
- 귀무가설( $H_0$ ) : 두 가지 교육방법에 따라 교육생의 만족율에 차이가 없다.

## <연구환경>

IT교육센터에서 PT를 이용한 프레젠테이션 교육방법과 실시간 코딩 교육 방법을 적용하여 교육을 실시하였다. 2가지 교육방법 중 더 효과적인 교육 방법을 조사하기 위해서 교육생 300명을 대상으로 설문을 실시하였다. 조사한 결과는 다음 표와 같다.

-----

```
# 대상 파일 : c:/Rwork/Part-III/two_sample.csv
# 해당 변수 : method(명목척도), survey(명목척도)
# 변수 척도 : 명목척도 : 빈도수(기술통계량)
```



# 두 집단 비율검정

<설문조사 교차표>

-----			
교육방법만족도	만족	불만족	참가자
-----			
PT교육	110	40	150
-----			
코딩교육	135	15	150
-----			
합계	245	55	300
-----			



# 두 집단 비율검정

## 1. 실습데이터 가져오기

```
getwd()
```

```
setwd("c:/Rwork/Part-III")
```

```
data <- read.csv("two_sample.csv", header=TRUE)
```

```
data
```

```
head(data) # 변수명 확인
```



# 두 집단 비율검정

## 2. 두 집단 subset 작성

`data$method # 1, 2 -> 노이즈 없음`

`data$survey # 1(만족), 0(불만족)`

`# 데이터 정제/전처리`

`x<- data$method # 교육방법(1, 2) -> 노이즈 없음`

`y<- data$survey # 만족도(1: 만족, 0:불만족)`

`x;y`



# 두 집단 비율검정

## 1) 데이터 확인

# 교육방법 1과 2 모두 150명 참여

table(x) # 1 : 150, 2 : 150

# 교육방법 만족/불만족

table(y) # 0 : 55, 1 : 245

## 2) data 전처리 & 기술통계량 -> 빈도수 -> 정규성 검정 필요 없음

# 두 변수에 대한 교차분석

**table(x, y, useNA="ifany") # 결측치 까지 출력**

#####

# y

#x 0 1

# 1 40 110 -> 방법A - 110 만족

# 2 15 135 -> 방법B - 135 만족

#####



# 두 집단 비율검정

## 3. 두 집단 비율차이검증 - prop.test()

`help(prop.test) # prop.test(x,n,p, alternative, conf.level, correct)`

**# 양측검정**

`prop.test(c(110,135),c(150,150)) # 방법A 만족도와 방법B 만족도 차이 검정`

**# p-value = 0.0003422**

`#sample estimates: 집단 간 비율`

`# prop 1 prop 2`

`#0.7333333 0.9000000`

`prop.test(c(110,135),c(150,150), alternative="two.sided", conf.level=0.95)`

**# 해설) p-value = 0.0003422 - 두 집단간의 만족도에 차이가 있다.**

**# 단측검정**

`prop.test(c(110,135),c(150,150), alter="greater", conf.level=0.95)`

**# 해설) p-value=0.9998 : 방법A가 방법B에 비해 만족도가 낮은 것으로 파악**



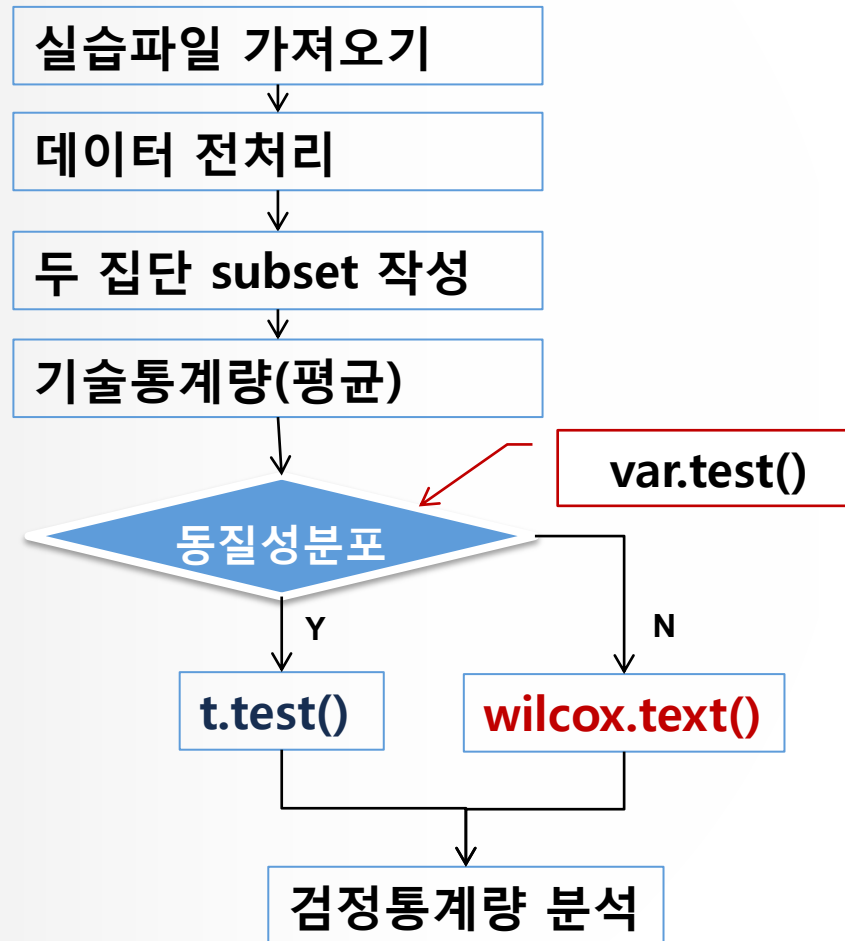
# 두 집단 평균검정

```
#####  
# 추론통계학 분석 - 2-2. 두 집단 평균 검정(독립표본 T검정)  
#####  
# 방법 : 두 집단 간 평균 차이에 관한 분석  
# 작업절차  
# 1. 실습파일 가져오기  
# 2. 두 집단 subset 작성(데이터 정제,전처리)  
# 3. 두 집단 간 동질성 검증(정규분포 검정)  
#     -> var.test()  
# 4. 두 집단 평균 차이검정  
#     -> t.test() or wilcox.test()  
#####
```



# 두 집단 평균검정

- 분석절차







# 두 집단 평균검정

## <연구가설>

- 연구가설( $H_1$ ) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다.
- 귀무가설( $H_0$ ) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 없다.

## <연구환경>

IT교육센터에서 PT를 이용한 프레젠테이션 교육방법과 실시간 코딩 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 150명을 대상으로 실기시험을 실시하였다. 두 집단간 실기시험의 평균에 차이가 있는가 검정한다.

-----

# 대상 파일 : c:/Rwork/Part-III/two\_sample.csv

# 해당 변수 : method(명목척도), score(비율척도)

# 대상 변수 : 교육방법, 시험성적

# 모형(모델) : 교육방법(A/B) -> 시험성적(비율-성적)



# 두 집단 평균검정

## 1. 실습파일 가져오기

```
data <- read.csv("c:/Rwork/Part-III/two_sample.csv", header=TRUE)
data
print(data)
head(data) #4개 변수 확인
summary(data) # score - NA's : 73개
```

## 2. 두 집단 subset 작성(데이터 정제, 전처리)

```
result <- subset(data, !is.na(score), c(method, score))
# c(method, score) : data의 전체 변수 중 두 변수만 추출
# !is.na(score) : na가 아닌 것만 추출
# 위에서 정제된 데이터를 대상으로 subset 생성
result # 방법1과 방법2 혼합됨
length(result$score) # 227
```



# 두 집단 평균검정

# 데이터 분리

1) 교육방법 별로 분리

```
a <- subset(result,method==1)
```

```
b <- subset(result,method==2)
```

2) 교육방법에서 점수 추출

```
a1 <- a$score
```

```
b1 <- b$score
```

# 기술통계량 -> 평균값 적용 -> 정규성 검정 필요

```
length(a1); # 109
```

```
length(b1); # 118
```



# 두 집단 평균검정

## 3. 분포모양 검정 : 두 집단의 분포모양 일치 여부 검정

# 귀무가설 : 두 집단 간 분포의 모양이 동질적이다.

# 두 집단간 동질성 비교(분포모양 분석)

**var.test(a1, b1) # p-value = 0.3002 -> 차이가 없다.**

# 동질성 분포 : t.test()

# 비동질성 분포 : wilcox.test()

## 4. 가설검정 - 두 집단 평균 차이검정

**t.test(a1, b1)**

**t.test(a1, b1, alter="two.sided", conf.int=TRUE, conf.level=0.95)**

# p-value = 0.0411 - 두 집단간 평균에 차이가 있다.

**t.test(a1, b1, alter="greater", conf.int=TRUE, conf.level=0.95)**

# p-value = 0.9794 : a1을 기준으로 비교 -> a1이 b1보다 크지 않다.

**t.test(a1, b1, alter="less", conf.int=TRUE, conf.level=0.95)**

# p-value = 0.02055 : a1이 b1보다 작다.



# 두 집단 평균검정

## 【독립표본 t-검정 결과 정리 및 기술】

1) 가설 설정	연구가설(H1) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다.
	귀무가설(H0) : 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다.
2) 연구환경	IT교육센터에서 PT를 이용한 프레젠테이션 교육방법과 실시간 코딩 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 150명을 대상으로 실기시험을 실시하였다. 두 집단간 실기시험의 평균에 차이가 있는가 검정한다.
3) 유의수준	$\alpha = 0.05$
4) 분석방법	독립표본 T검정
5) 검정통계량	$t = -2.0547, df = 218.192$
6) 유의확률	$P = 0.0411$
7) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 교육방법에 따른 두 집단 간 실기시험의 평균에 차이가 있다라고 말할 수 있다. 단측검정을 실시한 결과 교육방법1이 교육방법2보다 크지 않은 것으로 나타났다. 즉 실시간 코딩 교육방법이 교육효과가 더 높은 것으로 분석된다.



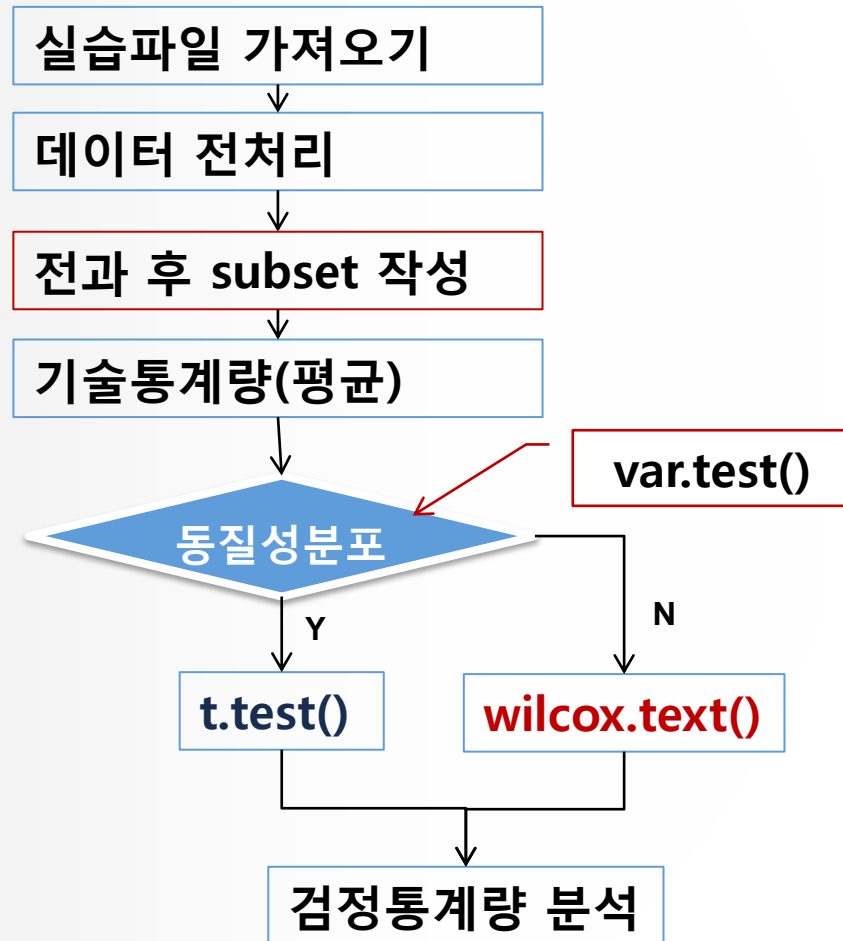
# 대응 두 집단 평균검정

```
#####  
# 추론통계학 분석 - 2-3. 대응 두 집단 평균 검정(대응표본 T검정)  
#####  
# 방법 : 대응되는 두 집단간 평균 차이에 관한 분석  
# 작업절차  
# 1. 실습파일 가져오기  
# 2. 두 집단 subset 작성(데이터 정제, 전처리)  
# 3. 두 집단 간 동질성 검증(정규분포 검정)  
# -> var.test(x,y paired=TRUE)  
# 4. 두 집단 평균 차이검정  
# -> t.test(x,y, paired=TRUE)  
# -> wilcox.test(x,y, paired=TRUE)  
#####
```



# 대응 두 집단 평균검정

- 분석절차





# 대응 두 집단 평균검정

## 1. 실습파일 가져오기

```
getwd()
```

```
setwd("c:/Rwork/Part-III")
```

```
data <- read.csv("paired_sample.csv", header=TRUE)
```

## 2. 두 집단 subset 작성

### 1) 데이터 정제

```
# subset(x, subset, select, ..) -> subset은 반드시 논리적이어야 함
```

```
result <- subset(data, !is.na(after), c(before,after))
```

```
# data 테이블을 대상으로 after 결측치 제거하여 subset 생성
```

```
result # 결측 데이터 4개
```





# 대응 두 집단 평균검정

## 2) 동일한 사람에게 두 번 질문

`x <- result$before` # 교수법 적용 전 점수

`y <- result$after` # 교수법 적용 후 점수

`x;y` # 대응포본인 경우 표본수가 같아야 한다. -> 짝을 이루어야 되기 때문에

`length(x)` # 96 -> 4개 결측치 제거

`length(y)` # 96

`mean(x)` # 5.16875

`mean(y)` # 6.220833 -> 1.052 정도 증가

## 3. 분포모양 검정 : 두 집단의 분포모양 일치 여부 검정

`var.test(x, y, paired=TRUE)` # p-value = 0.7361 -> 차이가 없다.

# 동질성 분포 : `t.test()`

# 비동질성 분포 : `wilcox.test()`



# 대응 두 집단 평균검정

## 4. 가설검정

```
t.test(x, y, paired=TRUE) # p-value < 2.2e-16
```

# 단측검정 - 방향성 검정

```
t.test(x, y, paired=TRUE, alter="greater", conf.int=TRUE, conf.level=0.95)
```

#p-value = 1 -> x을 기준으로 비교 : x가 y보다 크지 않다.

```
t.test(x, y, paired=TRUE, alter="less", conf.int=TRUE, conf.level=0.95)
```

# p-value < 2.2e-16 -> x을 기준으로 비교 : x가 y보다 적다.

### <해설>

교수법 프로그램을 적용하기 전 시험성적과 교수법 프로그램을 적용한 후 시험성적을 비교한 결과 교수법을 적용한 후 시험성적이 약 1.052 점수가 향상된 것으로 나타났다.



### 3) 대응표본 t-검정

#### 【대응표본 t-검정 결과 정리 및 기술】

1) 가설 설정	연구가설(H1) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 있다.
	귀무가설(H0) : 교수법 프로그램을 적용하기 전 학생들의 학습력과 교수법 프로그램을 적용한 후 학생들의 학습력에 차이가 없다.
2) 연구환경	A교육센터에서 교육생 100명을 대상으로 교수법 프로그램 적용 전에 실기시험을 실시한 후 1개월 동안 동일한 교육생에게 교수법 프로그램을 적용한 후 실기시험을 실시한 점수와 평균에 차이가 있는가 검정한다.
3) 유의수준	$\alpha = 0.05$
4) 분석방법	대응표본 T검정
5) 검정통계량	$t = -13.6424, df = 95$
6) 유의확률	$P = < 2.2e-16$
7) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 교수법 프로그램 적용 전과 적용 후의 두 집단 간 학습력의 평균에 차이가 있다. 라고 말할 수 있다. 또한 단측검정을 실시한 결과 교수법 프로그램 적용 전 학습력이 교수법 프로그램 적용 후 학습력 보다 크지 않은 것으로 나타났다. 즉 교수법 프로그램 이 학습력에 효과가 있는 것으로 분석된다.



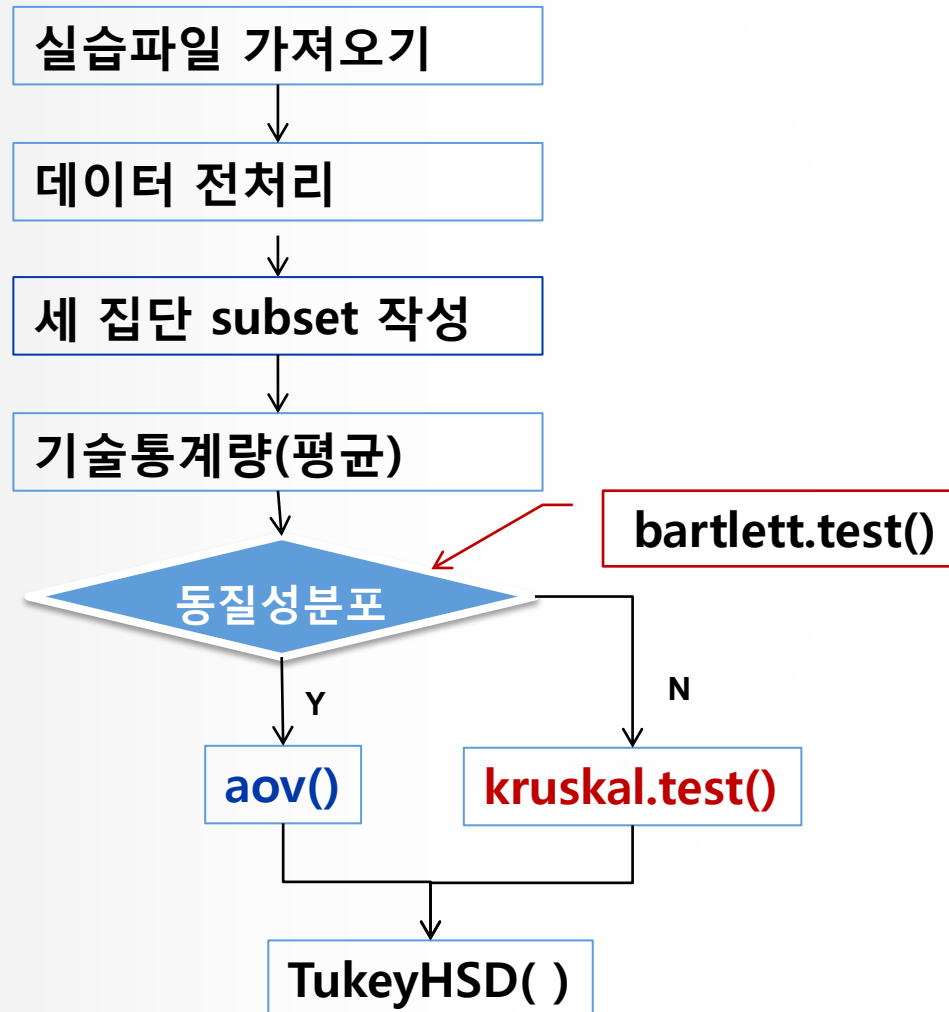
# 세 집단 평균 검정

```
#####  
# 추론통계학 분석 - 3. 세 집단 평균 검정(분산 분석)  
#####  
# 방법 : 세 집단(이상)간 평균 차이에 관한 분석  
# 작업절차  
# 1. 파일 가져오기  
# 2. 데이터 정제/전처리 - NA, outline 제거  
# 3. 세집단 subset 작성  
# -> 코딩 변경  
# -> 기술통계량(빈도수)  
# -> 교차표 작성  
# 4. 세집단 동질성 검정 : bartlett.test()  
# 5. 분산검정 : aov() or kruskal.test()  
# 6. 사후검정 : TukeyHSD()  
#####
```



# 세 집단 평균검정

- 분석절차





# 세 집단 평균검정

## <연구가설>

- 연구가설( $H_1$ ) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 있다.
- 귀무가설( $H_0$ ) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 없다.

## <연구환경>

세 가지 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 50명씩을 대상으로 실기시험을 실시하였다. 세 집단간 실기시험의 평균에 차이가 있는가 검정한다.

-----

- # 대상 파일 : c:/Rwork/Part-III/two\_sample.csv
- # 해당 변수 : method(명목척도), score(비율척도)
- # 대상 변수 : 교육방법, 시험성적
- # 모형(모델) : 교육방법(A/B) -> 시험성적(비율-성적)



# 세 집단 평균 검정

## 1. 파일 가져오기

```
data <- read.csv("c:/Rwork/Part-III/three_sample.csv", header=TRUE)
```

## 2. 데이터 정제/전처리 - NA, outline 제거

```
data <- subset(data, !is.na(score), c(method, score))
```

```
data # method, score
```

```
# 차트이용 - online 보기(데이터 분포 현황 분석)
```

```
plot(data$score) # 차트로 outline 확인 : 50이상과 음수값
```

```
barplot(data$score) # 바 차트
```

```
boxplot(data$score) # 박스 차트
```

```
mean(data$score) # 14.45
```



# 세 집단 평균 검정

```
# outline 제거 - 평균(14) 이상 제거
length(data$score)#91
data2 <- subset(data, score <= 14) # 14이상 제거
length(data2$score) #88(3개 제거)

##### 정제된 데이터 보기 #####
x <- data2$score
boxplot(x)
plot(x)
bp <- boxplot(data2$score) # 차트 결과 저장
```





# 세 집단 평균 검정

## 3. 세 집단 subset 작성

# 코딩 변경 - 변수 리코딩 -> method: 1:방법1, 2:방법2, 3:방법3

**data2\$method2[data2\$method==1] <- "방법1"**

**data2\$method2[data2\$method==2] <- "방법2"**

**data2\$method2[data2\$method==3] <- "방법3"**

**table(data2\$method2) # 교육방법 별 빈도수**

#방법1 방법2 방법3

# 31 27 30

**x <- table(data2\$method2)**

#교육방법에 따른 시험성적 평균 구하기

**y <- tapply(data2\$score, data2\$method2, mean)**

# 방법1 방법2 방법3

# 4.187097 6.800000 5.610000

**out <- data.frame(교육방법=x, 시험성적=y)**

**out # 교육방법에 따른 시험성적 평균 교차표**

# 교육방법.Var1 교육방법.Freq 시험성적

#방법1 방법1 31 4.187097

#방법2 방법2 27 6.800000

#방법3 방법3 30 5.610000



# 세 집단 평균 검정

## 4. 동질성 검정 - 정규성 검정

```
# bartlett.test(종속변수 ~ 독립변수) # 독립변수 - 세 집단
bartlett.test(score ~ method, data=data2)
#Bartlett's K-squared = 3.3157, df = 2, p-value = 0.1905
```

**# data2의 테이블을 대상으로**

# 3집단 이상인 경우 : (종속변수 ~ 독립변수) 분석식으로 표현

# ~ : 틸드 -> 집단별로 subset를 만들지 않고 사용하도록 편의성 제공

# 귀무가설 : 세 집단 간 분포의 모양이 동질적이다.

# 해설 : 유의수준 크기 때문에 귀무가설을 기각할 수 없다.

# 동질한 경우 aov() 사용 : aov - Analysis of Variance(분산분석)

# 동질하지 않은 경우 - kruskal.test()



# 세 집단 평균 검정

## 5. 분산검정

**help(aov)**

# 분산분석 결과를 result에 저장

# 귀무가설 : 세 집단의 평균에 차이가 없다.

**data2\$method2 <- factor(data2\$method2)**

# factor() : method가 집단 구성변수라는 것을 명시

# aov(종속변수 ~ 독립변수, data=data set)

**result <- aov(score ~ method2, data=data2)**

**names(result)**

# aov()의 결과값은 summary()함수를 사용해야 p-value 확인

**summary(result) # Pr(>F) : 9.39e-14 -> 귀무가설 기각**

# 해설 : 0.05보다 현저하게 작음

# 교육방법에 따라서 시험성적 평균에 차이가 있다.



# 세 집단 평균 검정

## 6. 사후검정

# 집단간 차이 상세보기 ->  $A \neq B \neq C$ ,  $A = B \neq C$ ,  $A \neq B = C$

**TukeyHSD(result) # 분석분석의 결과로 사후검정**

# \$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
----------	----------	-----------	-----------	-----------

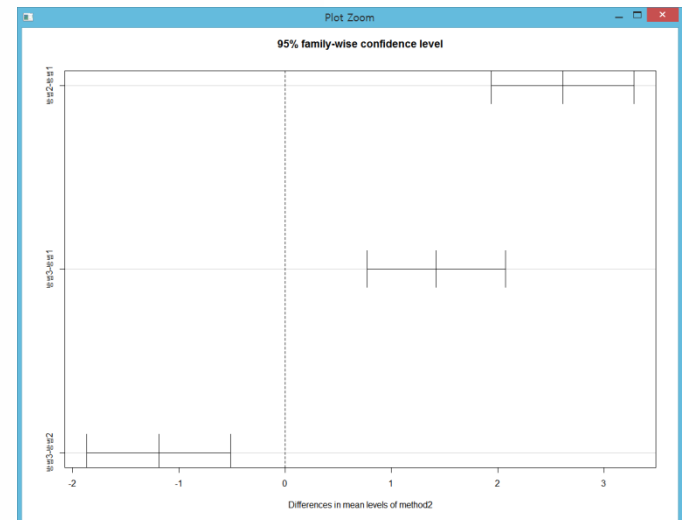
#방법3-방법2	-1.190000	-1.8656509	-0.5143491	0.0001911
----------	-----------	------------	------------	-----------

# 교육방법 간 비교 -> p값(tapply 차이 검정) -> 4.187097 6.800000 5.610000

# 해석) A B C 집단간 모두 차이가 있다.

**plot(TukeyHSD(result))**

# 그래프 보기(lwr~upr변수 이용)





# 세 집단 평균 검정

## 【분산분석 결과 정리 및 기술】

1) 가설 설정	연구가설(H1) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 있다.
	귀무가설(H0) : 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 없다.
2) 연구환경	세 가지 교육방법을 적용하여 1개월 동안 교육받은 교육생 각 50명씩을 대상으로 실기시험을 실시하였다. 세 집단간 실기시험의 평균에 차이가 있는가 검정한다.
3) 유의수준	$\alpha = 0.05$
4) 분석방법	ANOVA 검정
5) 검정통계량	<b><math>F = 43.58, Df = 2, Sum Sq = 99.37, Mean Sq = 49.68</math></b>
6) 유의확률	<b><math>P = 9.39e-14 ***</math></b>
7) 결과해석	유의수준 0.05에서 귀무가설이 기각되었다. 따라서 교육방법에 따른 세 집단 간 실기시험의 평균에 차이가 있는 것으로 나타났다. 또한 사후검정 방법인 Tukey 분석을 실시한 결과 '방법2-방법1'의 평균 점수의 차이가 가장 높은 것으로 나타났다.



# 14. 요인분석과 상관분석

## Chap14\_1\_FactorAnalysis 수업내용

- 1) 요인분석 개요
- 2) 공통요인으로 변수 정제



# 1) 요인분석 개요

## 요인분석(Factor Analysis)

- 다수의 변수들을 대상으로 변수들 간의 관계 분석
- 공통 차원으로 축약하는 통계기법
- 탐색적 요인분석과 확인적 요인분석으로 구분
- 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 조사하는 방법



# 1) 요인분석 개요

## 【요인분석의 전제조건】

- 하위요인으로 구성되는 데이터 셋이 준비되어 있어야 한다.
- 분석에 사용되는 변수는 등간척도나 비율척도이어야 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.





# 1) 요인분석 개요

## 【요인분석의 목적】

- 자료의 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 용인으로 묶이는지 여부를 확인



## 2) 공통요인으로 변수 정제

### 【데이터 셋 준비】

# name : 각 과목의 문제에 대한 문항 이름

```
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)
```

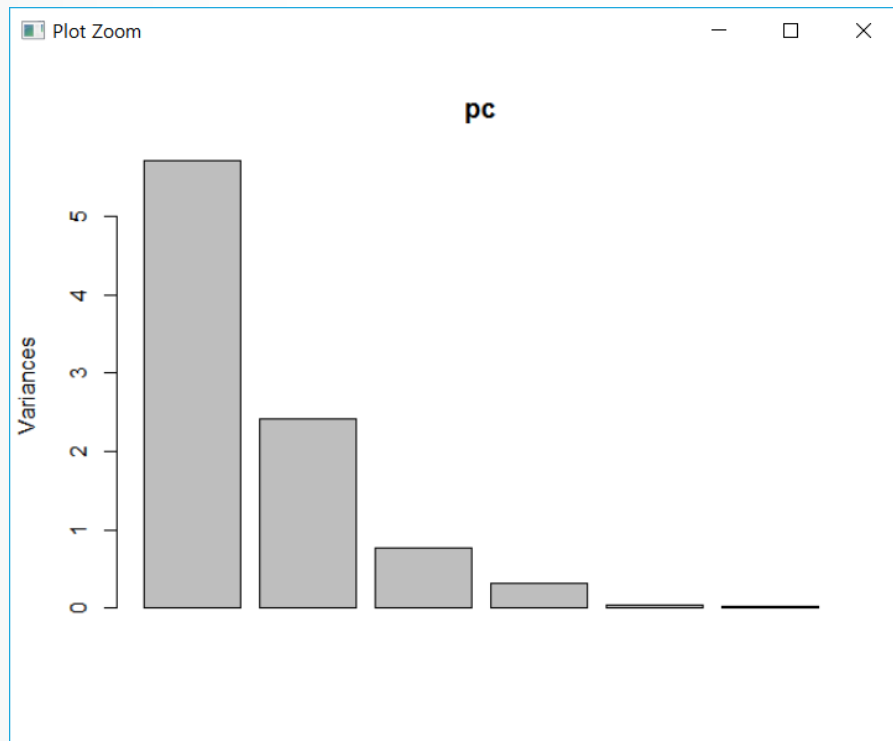
```
name <- 1:10
```



## 2) 공통요인으로 변수 정제

### 【주성분분석 요인 수 분석】

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

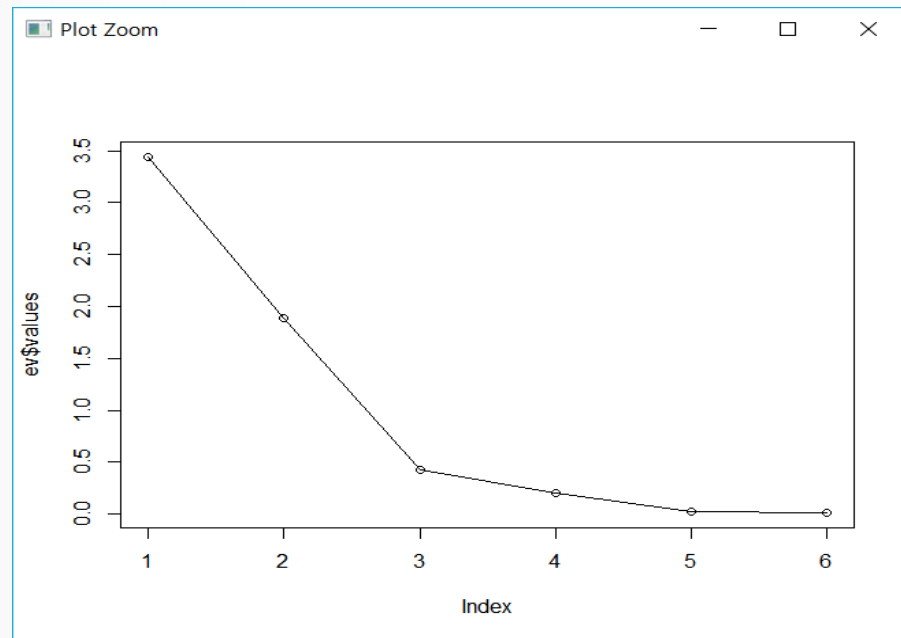




## 2) 공통요인으로 변수 정제

### 【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")
```





## 2) 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")
```

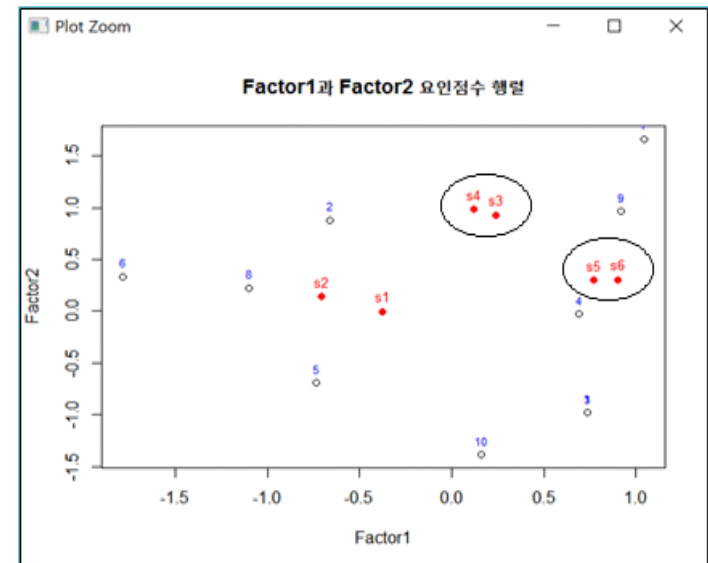
```
# 요인적재량의 레이블 표시
```

```
text(result$loadings[,1], result$loadings[,2],
```

```
labels =
```

```
rownames(result$loadings),
```

```
cex = 0.8, pos = 3, col = "red")
```





## 2) 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")
```

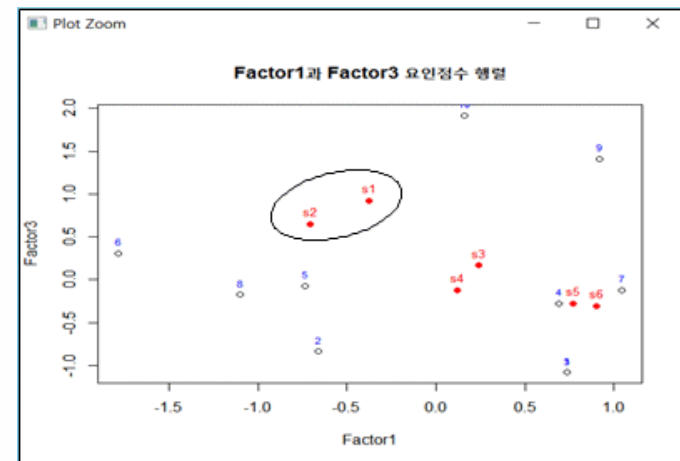
```
# 요인적재량 추가
```

```
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")
```





# 14. 요인분석과 상관분석

## Chap14\_2\_CorrelationAnalysis 수업내용

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



# 1) 상관분석 개요

## 상관관계 분석(Correlation Analysis)

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





## 2) 피어슨 상관계수 $r$

### 【피어슨 상관계수 $R$ 】

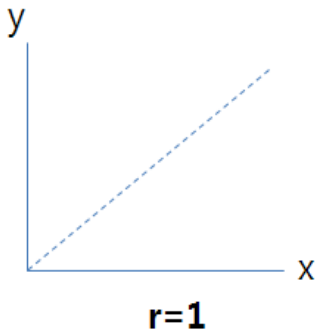
피어슨 상관계수 $R$	상관관계 정도
$\pm 0.9$ 이상	매우 높은 상관관계
$\pm 0.9 \sim \pm 0.7$	높은 상관관계
$\pm 0.7 \sim \pm 0.4$	다소 높은 상관관계
$\pm 0.4 \sim \pm 0.2$	낮은 상관관계
$\pm 0.2$ 미만	상관관계 없음
※ 상관계수 $r$ 은 -1에서 +1까지의 값을 가진다. 또한 가장 높은 완전 상관관계의 상관계수는 1이고, 두 변수간에 전혀 상관관계가 없으면 상관계수는 0이다.	



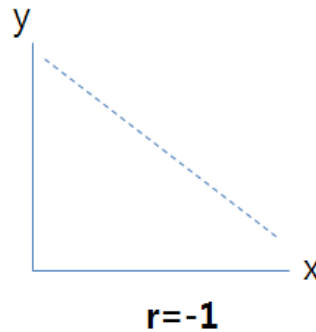
## 2) 피어슨 상관계수 $r$

- 상관계수  $r$ 과 상관관계 정도

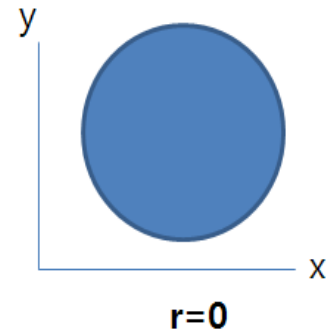
완전 정(+) 상관관계



완전 부(-) 상관관계



상관관계 없음





### 3) 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)
```



## 4) 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