연구데이터 분석 기본과정

2015. 04. 16

㈜ 아이티베인 이 현 우

[연구 데이터 분석 기본과정]

제1장 기초 통계

회사 직원들의 직업관과 사내생활에 대한 만족도를 조사하여 전략적인 커뮤니케이션과 효율적인 인사관리의 자료를 얻기 위함

회사에 대한 평가 : Q1
조직에 대한 신뢰 및 존중 : Q2, Q3, Q4
업무지원 : Q5
업무에 대한 흥미도 : Q6
기회의 제공 : Q7

각각의 개념은 몇 개의 소문항으로 구성되어 있음

엔진 제어 모듈에서 쓰이는 미세한 디바이스는 리드(lead) 사이의 거리가 650micron(100만분의 1m)이다. 이 리드는 디바이스가 외부와 ‘연락’ 할 수 있게 해주는 작은 선들이다. 로봇 기계는 이 디바이스를 집어서 회로판에 갖다 놓는 역할을 한다.

조사의 일환으로 특정한 형태의 미세한 디바이스가 서로 다른 네 가지 속도로 회로판에 놓여지고 이러한 시행이 각 속도별로 16번 측정하여 한쪽 방향으로 치우침 정도의 결과값이다.

기계 속도와 치우침의 정도 사이에 관계가 있는가?

기 계 속 도							
1		2		3		4	
0.0639	0.0744	0.0808	0.0479	0.0737	0.0936	0.0476	0.0591
0.0755	0.0720	0.0704	0.0737	0.0632	0.0756	0.0640	0.0451
0.0595	0.0698	0.0632	0.0803	0.0784	0.0815	0.0511	0.0633
0.0846	0.0530	0.0846	0.0711	0.0806	0.0893	0.0559	0.0202
0.0533	0.0690	0.0741	0.0552	0.0912	0.0864	0.0785	0.0423
0.0637	0.0558	0.0591	0.0707	0.0711	0.0794	0.0392	0.0463
0.0673	0.0713	0.0500	0.0584	0.0915	0.0643	0.0469	0.0463
0.0781	0.0715	0.0772	0.0791	0.0733	0.0512	0.0549	0.0350

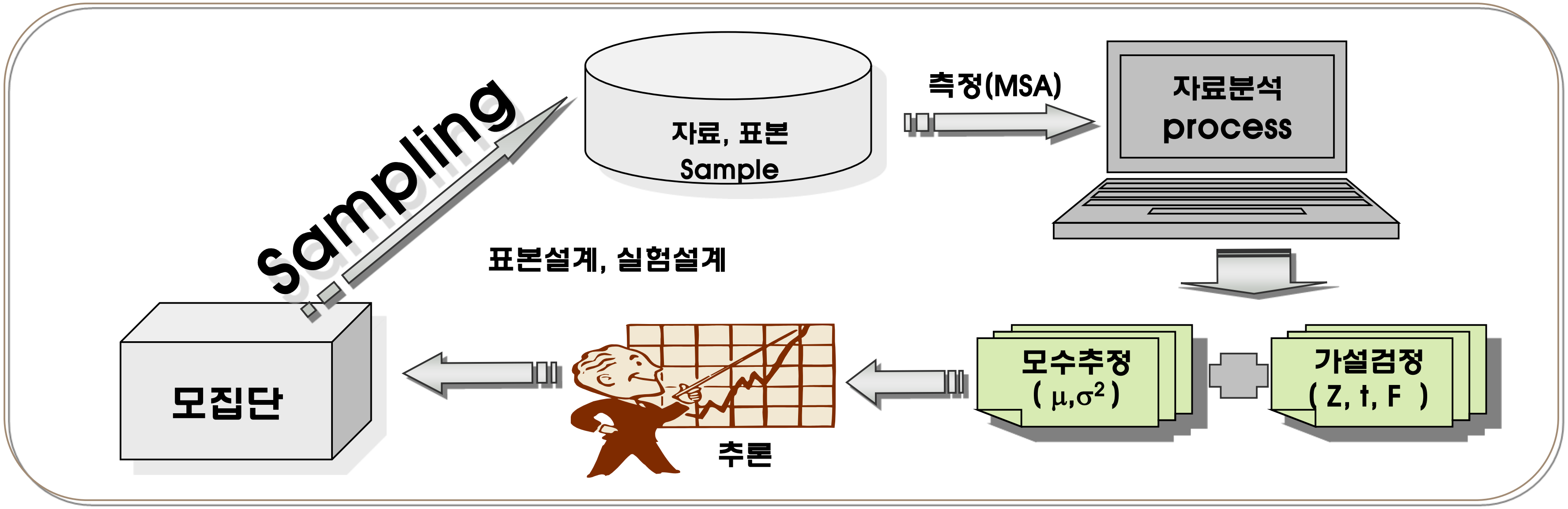
다음 Data는 어떤 기계의 사용 빈도와 그 기계의 수리비용이다. 사용빈도가 기계의 수리비용에 영향을 주는가?

x	y	x	y	x	y
531	22.99	529	23.01	533	23.14
535	23.36	535	23.42	535	23.11
536	23.62	534	23.16	530	23.24
530	22.86	526	22.87	531	23.13
532	23.16	533	23.62	530	23.00
533	23.28	534	23.63	531	23.35
532	22.89	530	23.01	529	22.62
531	23.00	531	23.12	534	23.37
528	23.08	536	23.50	532	23.08
534	23.64	533	22.75	533	23.31

통계학(Statistics)의 정의

State + Arithmetic (국가 + 산술) 의 의미로 시작

- 관심의 대상에 대한 자료를 수집,
- 자료에 여러 가지 통계적 기법을 적용하여 정보를 추출,
- 관심의 대상에 대한 특성을 파악,
- 의사결정을 지원해 주는 학문



통계학(Statistics)의 정의

『 불확실하에서 현명한 의사 결정을 위해 필요한 **자료를 수집, 분석**하여 유일한 **수자적 정보를 제시**하고 **통계적 법측을 발견**하는 이론과 방법을 연구하는 지식체계 』 *–" 통계학의 정의 " –*

통계의 어원

Statistic(단수)

– 평균치,지수, 표준편차, 상관계수 등과 같이 통계집단의 특성치

Statistics(복수)

– 통계자료

Status 라틴어의 "상태"

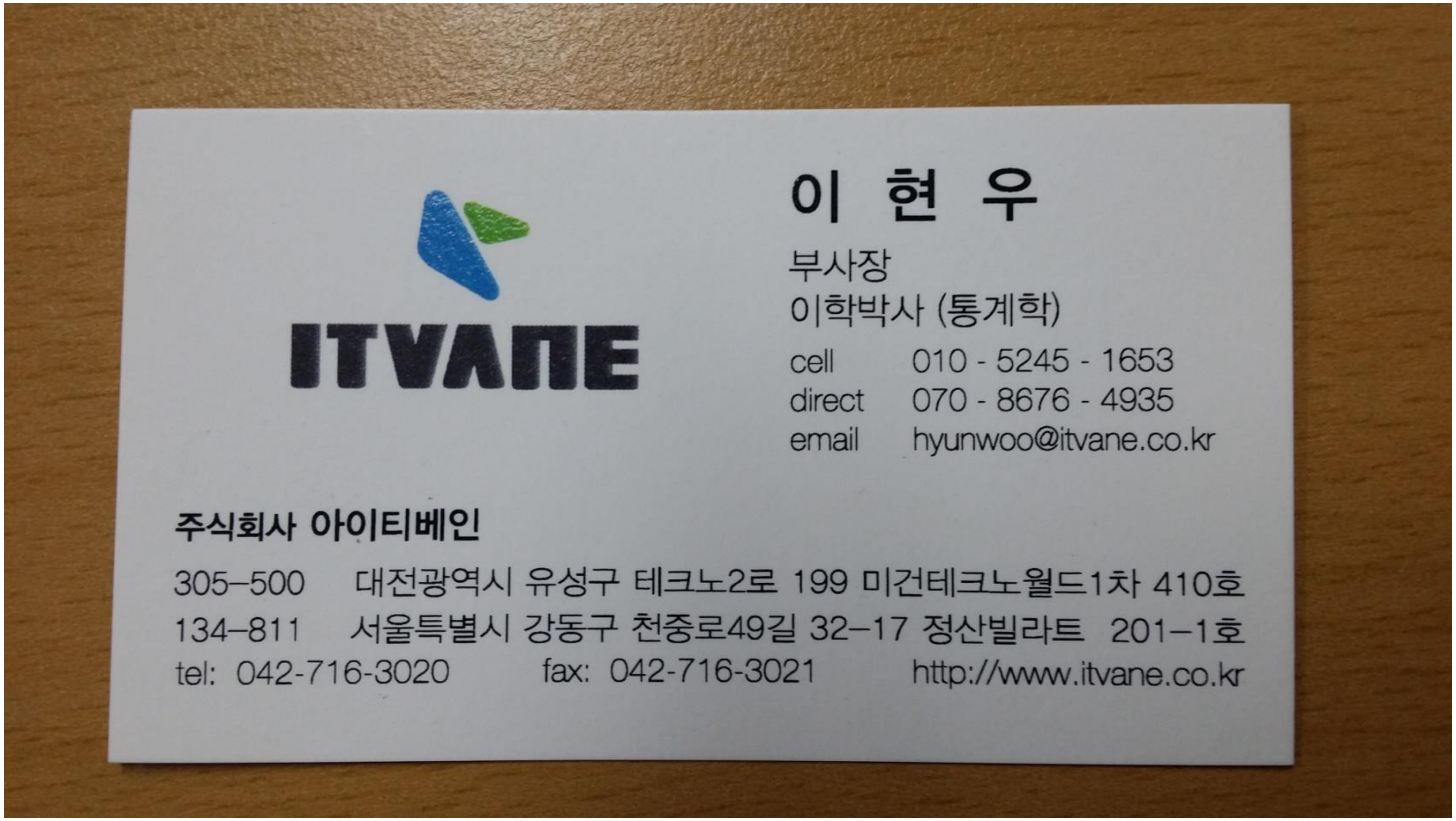
\+

Statista 이탈리아의 "정치학"

" State arithmetic(국가산출)

역사적으로 정치가들이 국가의 살림을 꾸려 나가기 위하여 필요한 숫자를 체계적으로 산출해 내는 데서 유래 "

1.1 통계의 이해



ITVANE
이 현 우
부사장
이학박사 (통계학)
cell 010 - 5245 - 1653
direct 070 - 8676 - 4935
email hyunwoo@itvane.co.kr
주식회사 아이티베인
305–500 대전광역시 유성구 테크노2로 199 미건테크노월드1차 410호
134–811 서울특별시 강동구 천중로49길 32–17 정산빌라트 201–1호
tel: 042-716-3020 fax: 042-716-3021 http://www.itvane.co.kr

1.1 통계의 이해

- **얼마나 많이 숫자와 접하고 있나?**
- **숫자를 써서 공격하라.**
  - 영국의 수상 벤저민 디즈레일리(1804~1881)
  - There are three types of lies - lies, damn lies, and statistics.
- **우리나라 사람들이 숫자에 약한 이유**
- **통계 : 관심의 대상을 정리, 숫자로 표현한 것**

통계를 잘못 사용하고 있는 사례 : 여론조사

- 전수조사, 표본조사
- 표본조사방법 : 우편, 면접, 전화, 인터넷
- 장님 코끼리 만지기
- 1936년 미국 대통령 선거

  공화당의 랜던, 민주당의 루즈벨트

  Literary Digest

  1000만명의 유권자에게 설문지 우송, 230만명에게 회신

  결과 : 랜던의 여유있는 승리

  가장 유명한 실수

  원인 : 잡지의 정기 구독자, 전화번호부

통계를 잘못 사용하고 있는 사례 : 너무 정확한 통계

- 오스트리아 재무부의 공식 출판물

  1951년도 잘츠부르크 인구가 전체인구의

  4.719303%

- 로치(Hal Roach)라는 코메디언 – 자연사 박물관
- 벽제의 공동묘지를 다녀간 인원

  12시 까지 – 7,865명, 이후 – 2,376명

- 너무 정확한 표현은?

통계를 잘못 사용하고 있는 사례 : 잘못된 해석

- 미국의 한 조사 발표

교회에 다니는 사람들은 결혼 생활을 계속 유지한다.

이혼 소송중인 95%가 부부 중 한 사람 혹은 둘 다 교회에 정기적으로 가지 않는다.

이혼소송중인 부부/ 결혼 생활을 유지하는 부부

- 모집단의 크기 문제

	충청권	비 충청권	전체	찬성율
모집단 크기	10,000	90,000	100,000	
응답자 수	2,000	3,000	5,000	
찬성	1,800	900	2,700	54%
실제	9,000	27,000	36,000	36%

통계를 잘못 사용하고 있는 사례 – 매개변수

- 미국의 껌판매량과 범죄수의 관계
- 교회의 수가 늘어나면 범죄 발생률도 증가?
- 우유를 많이 마시면 암에 걸릴 확률이 증가

  우유를 많이 소비하는 미국의 북부, 중부 남부

  많이 마시지 않는 스리랑카

  우유를 많이 마시는 영국여자가 일본 여자들보다 18배나 더 많이 암에 걸린다.

  첫 번째 : 수명이 길다. 노년층이 많다.

  두 번째, 영국여자의 평균수명이 일본여자보다 12세 길다.
- 미국 메사추세츄의 장로교 목사의 월급과 쿠바 하바나의 럼주 가격간에는 높은 상관관계
- 우리나라 냉장고의 보급률과 위암환자의 수는 큰 상관관계

심프슨의 파라독스

경증

항암제	생존	사망	합계	생존률
New	**18**	**12**	**30**	**60%**
Old	**7**	**3**	**10**	**70%**
전체	**25**	**15**	**40**	

중증

항암제	생존	사망	합계	생존률
New	**2**	**8**	**10**	**20%**
Old	**9**	**21**	**30**	**30%**
전체	**11**	**29**	**40**	

실제

항암제	생존	사망	합계	생존률
New	20	20	40	50%
Old	16	24	40	40%
전체	36	44	80	

확률의 의미

- 확률의 의미

  P(A) : A라는 사상이 일어날 확률 ?

  A : 동전을 던졌을 때 앞면, 비가 온다

- 야구타율 : 3할
- 어느 의사 – 수술성공률 1%
- 딸만 일곱 낳은 사연
- %와 % 포인트
- 평균, 중앙값, 최빈수
- 1994년 미 프로야구 파업
  - 구단주 : 평균연봉 9억원
  - CBS의 여론조사 : 구단주 지지 43%, 선수 22%
  - 700여명의 메이저 리그의 평균연봉 : 9억원
  - 중앙값 : 3억원 , 최빈수 : 2억여원

통계학의 분류

통계학의 과거와 현재

기술통계 : 자료를 표와 그림으로 표현

통계적 의사결정 단계(추론통계)

자료를 통하여 모집단에 대한 어떤 특성을

일반화 하는 방법

1) 데이터의 중요성

데이터의 수집과 정리

Garbage in, garbage out !

- 연구와 분석의 목적을 명확히 해야 한다.
- 분석의 목적에 부합하는 데이터를 수집해야 한다.
- 데이터는 정밀하게 검사되고 분석에 적합하도록 정리되어야 한다.

1) 데이터의 중요성

오류값(Error) : 변수가 가질 수 없는 값, 변수값의 불가능한 조합, 일관성 없는 코드값, 잘못된 코드값.

특이값(Outlier) : 정상이 아닌 자료값. 특이값은 오류값일 수도 있고 그렇지 않을 수도 있다.

결측값(Missing) : 원인과 기록방법을 정밀하게 조사하여 자료를 정정하고 기록방법을 변경해야 하며, 필요 시에는 자료를 보정해야 한다.

사례	x1	x2	x3	x4	x5
1	76.7	Good	9	2.06	7.7
2	73.6	Good	7	2.14	7.4
3	68.7	Bad	3	4.21	6.9
4	9999	Reject	NA	.	0
5	82.7	Good	9	2.00	0.8
6	73.4	Bad	10	1.34	7.3
7	.	Good	2	2.20	0
8	69.5	Good	7	2.37	7.0
9	.	Good	3	1.82	0
10	69.5	Good	7	23.7	7.0

2) 분석방법

기술통계학 (Descriptive Statistics)

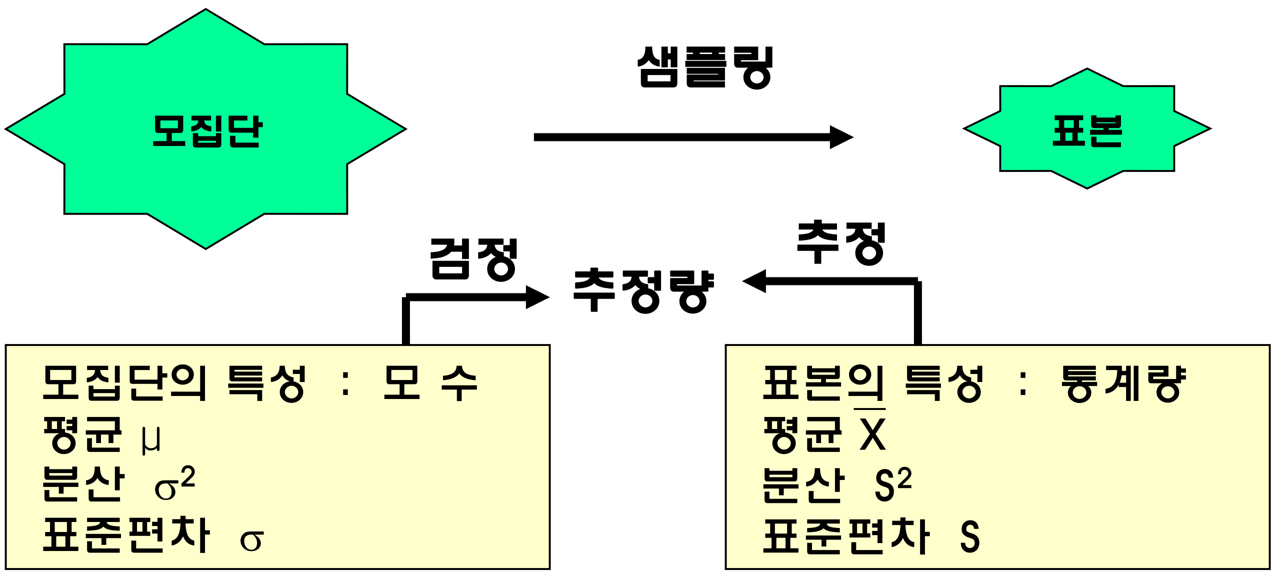
방대한 자료를 그래프나 몇 개의 숫자로 요약하여, 그 자료의 전반적인 내용을 쉽고 빠르게 파악할 수 있는 기법을 다루는 통계학.

추측통계학 (Inferential Statistics)

관심의 대상이 되는 전체집단(모집단)으로부터 모집단의 일부를 추출하여 관측된(표본) 내용을 근거로 하여 모집단의 전체 특성을 추측하고 검정(추론)하는 통계적 방법을 다루는 통계학

2) 분석방법

관심의 대상이 되는 모든 개체의 집합을 모집단 이라고 하며,
모집단에서 조사대상으로 채택된 일부를 표본이라고 한다.



모집단의 모수를 정확히 계산할 수 있다면 문제가 없으나,
이를 알기 어려운 상황에서는 표본에서 계산된 통계량을 바탕으로 모수를 추정한다.

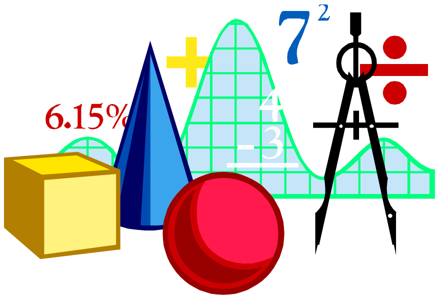
1.2 통계학과 자료분석

Data의 구분

• 정량적 특성(Quantitative Characteristic)

크기를 수치로 나타낼 수 있는 특성

1) 이산특성(Discrete Characteristic): 불연속적인 특성
   예: 공정상의 결점 수, 부적합 수, 고객불만 건수등
2) 연속특성(Continuous Characteristic): 연속적인 특성
   예: 제품 두께, 반사율, 점도, 밀도, 제품 강도(Strength)등



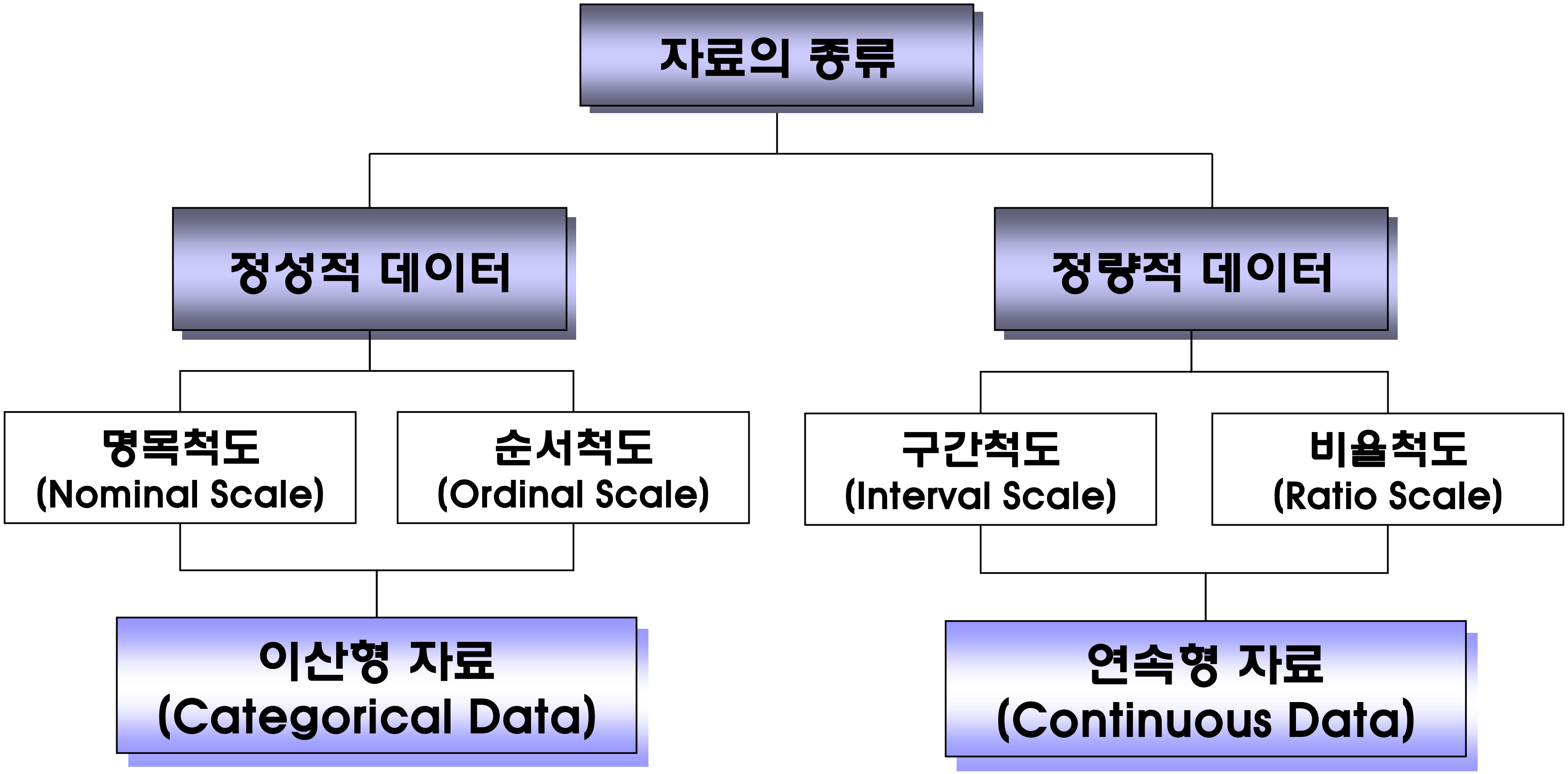
• 정성적 특성(Qualitative Characteristic)

크기를 특성(Attribute)으로 나타낼 수 있는 특성

1) 분류특성(Classified Attribute): 여러가지로 구별되는 특성
   예: 제품 Type, 제품 색상, 제품 등급등
2) 양자특성(Go/No-go Attribute): 두가지로 나눠지는 특성
   예: 합격/불합격, 양품/불량등

1.2 통계학과 자료분석

자료의 종류



이산형 자료

- **명목척도(Nominal Scale)**
  - 어떤 범주에 대해 단지 명목상 수치를 부여한 척도

    예) 성별 : 남자=1, 여자=2

    이뇨제의 종류 : 다이아자이드, 라식스, 알닥톤, 로졸
  - ***빈도분석, 교차분석, 원도표, 막대도표 범주형 데이터 분석***
- **순서척도(Ordinal Scale)**
  - 범주에 대해 속성의 순서에 따라 수치를 부여한 척도

    예) 건강상태 : 양호=3, 보통=2, 나쁨=1

    각종 점수

    학력 : 초등졸이하=1, 중졸=2, 고졸=3, 대졸=4, 대학원이상=5
  - ***빈도분석, 교차분석, 범주형 자료분석, 다변량 분석***

연속형 자료

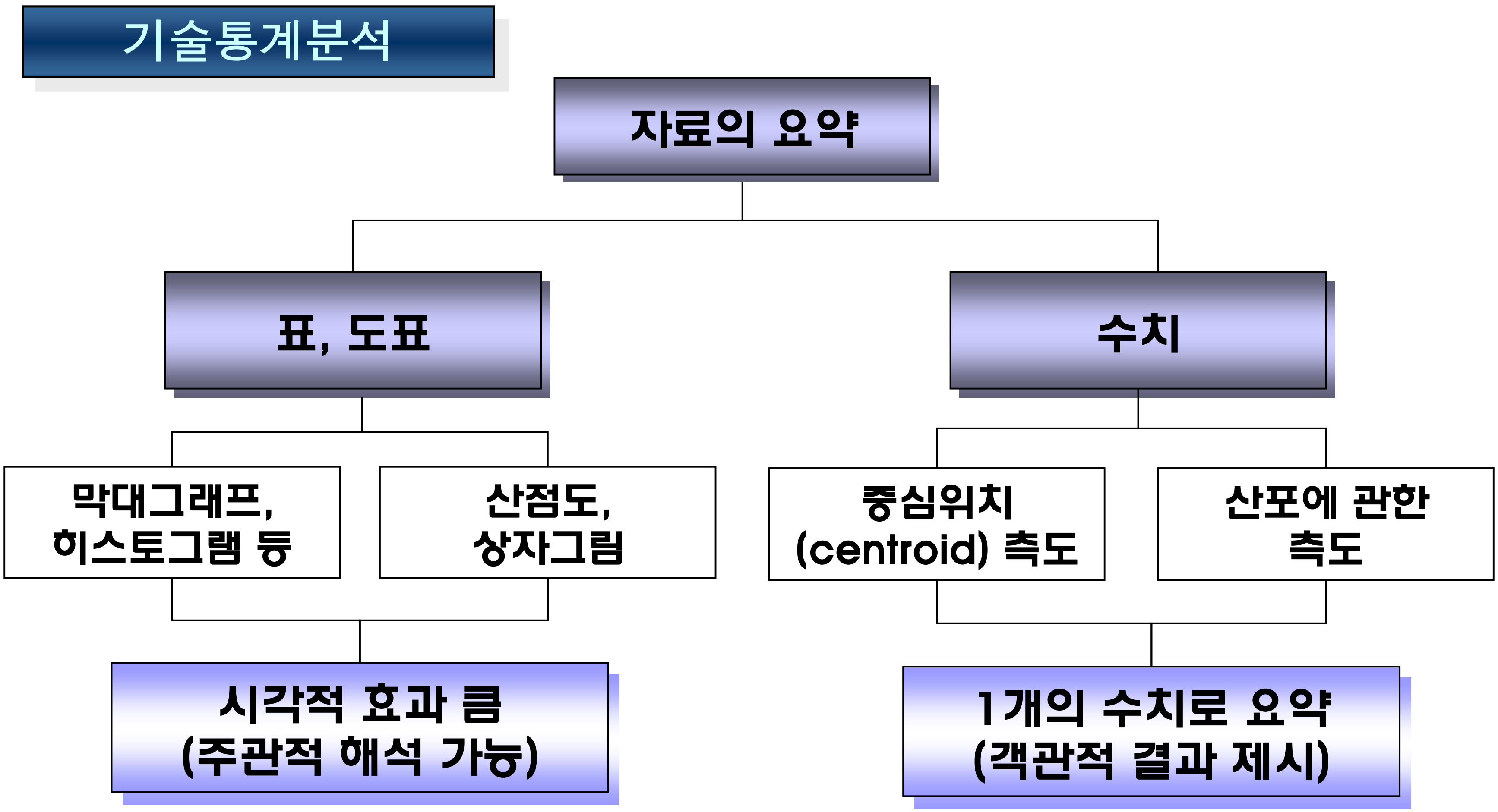
- **구간척도(Interval Scale)**
  - 절대 '영'(Absolute zero)이 없으며, 대상이 갖는 양적인 정도의 차이에 따라 등간격으로 수치를 부여한 척도

    예) 온도 : 섭씨 0℃, 50℃, 100℃

    물가지수, 산업생산지수, 무역수지 등
  - 수학적 의미 : (A−B)+(B−C)=A−C,
  - 표현 : 온도차, 물가지수 상승, 두 배로 덥다?
  - ***기술통계, 집단간 평균비교, 회귀분석, 다변량 분석***

- **비율척도(Ratio Scale)**
  - 절대 '영'이 존재하며, 비율계산이 가능한 수치를 부여한 척도

    예) 광고비, 판매량, 매출액, 무게, 가격, 소득 등
  - 수학적 의미 : 사칙연산이 가능함
  - ***기술통계, 집단간 평균비교, 회귀분석, 다변량 분석***

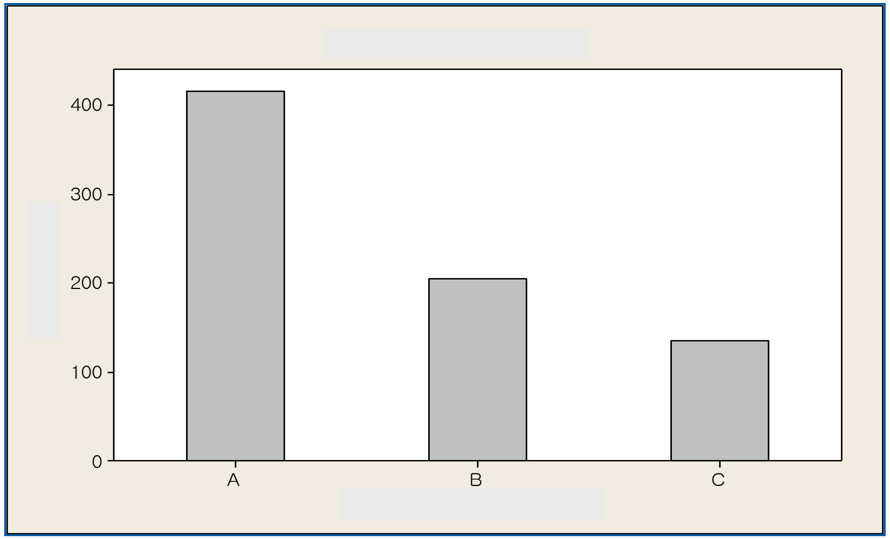

기술통계분석
자료의 요약
표, 도표
수치
막대그래프, 히스토그램 등
산점도, 상자그림
중심위치 (centroid) 측도
산포에 관한 측도
시각적 효과 큼 (주관적 해석 가능)
1개의 수치로 요약 (객관적 결과 제시)

1) 자료의 시각적 정리

❑ 막대그림 (Bar Chart)

- 이산형 자료일 경우 각 자료 값의 도수 (또는 상대도수)를 같은 폭의 막대로 표현한 그림
- 수평축은 일정한 폭을 지닌 수직막대를 통해 비교할 항목을 나열
- 수직 축은 막대의 높이(자료 값의 도수)에 의해 양을 표시

[막대그림]

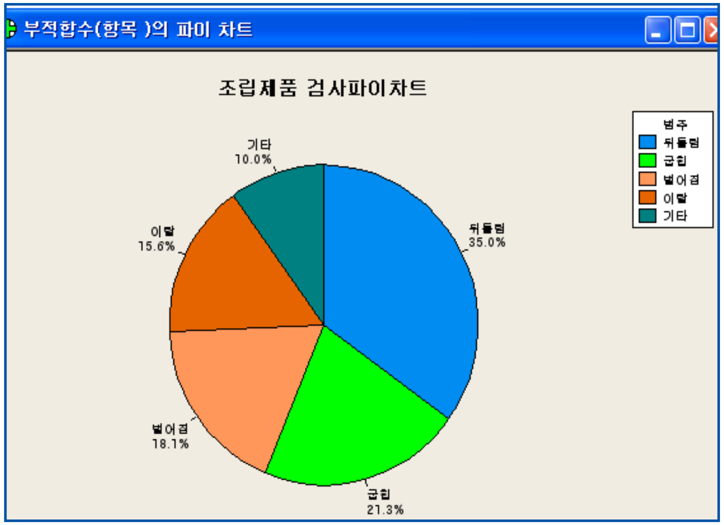


1) 자료의 시각적 정리

❑ 원그림 (Pie Chart)

- 원을 자료 값의 상대도수에 비례하도록 조각으로 나누어 표현한 그림
- 전체에 있어서 각 항목들의 상대적인 점유량을 표시
- 신문이나 잡지에서 많이 사용하는 그림
- 도수 설정, 구간조정 가능, 정리된 자료도 표현 가능

[파이 차트]

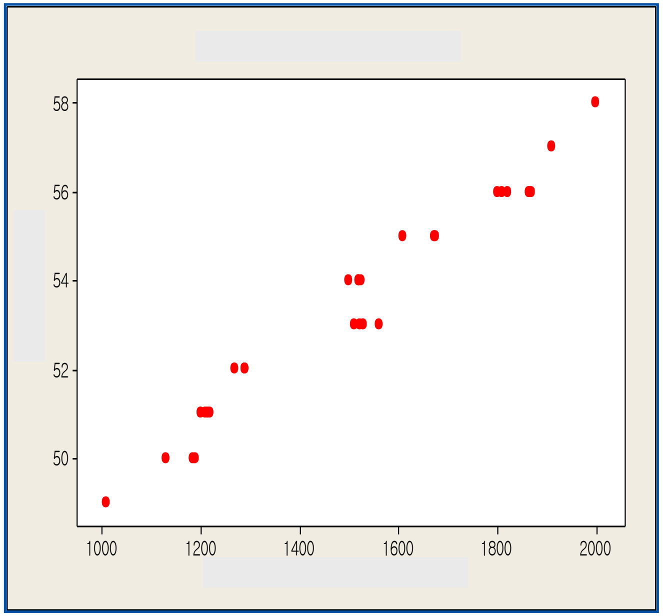


1) 자료의 시각적 정리

❑ 산점도 (Scatter Plot)

- 두 연속형 자료에 대하여 X축, Y축으로 하여 좌표값을 점으로 표시
- 두 연속형 자료의 관계를 분석하는데 매우 효율적

[산 점 도(S)]

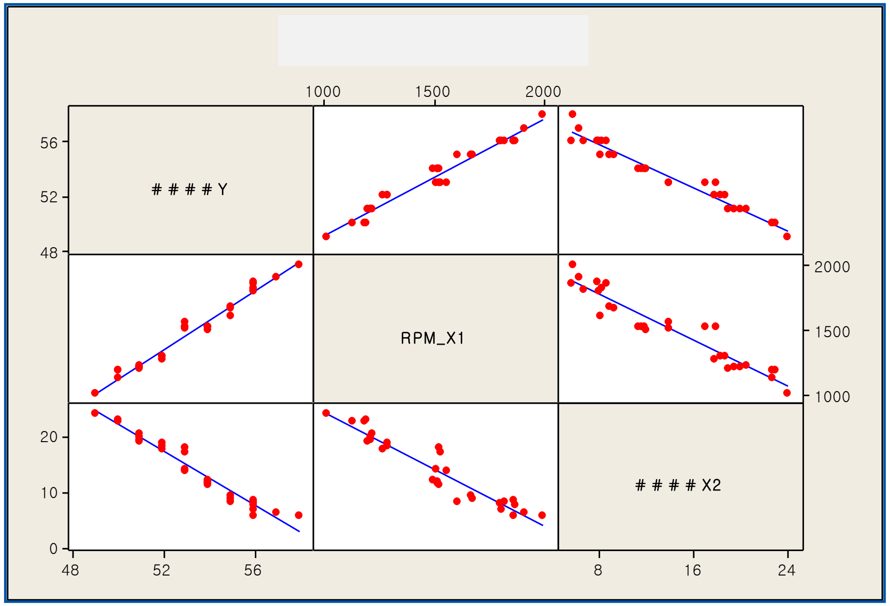


1) 자료의 시각적 정리

❑ 산점도 행렬 (Scatter Plot Matrix)

- 여러 개의 변수에 대하여 산점도를 동시에 그려주는 그림
- 같은 변수의 해당그림은 산점도 대신 히스토그램으로 표현

[행 렬 도(M)]



1) 자료의 시각적 정리

- **줄기 잎 그림 (Stem –and –Leaf Plot)**
  - **Raw Data의 정보를 그대로 유지하면서 관측값의 범위, 분포형태, 집중도 등의 전반적인 분포형태를 보여 준다.**
  - **Data 수가 많으면 오히려 분포의 형태를 파악하기가 어렵다.**

[줄기–잎–그림(F)]

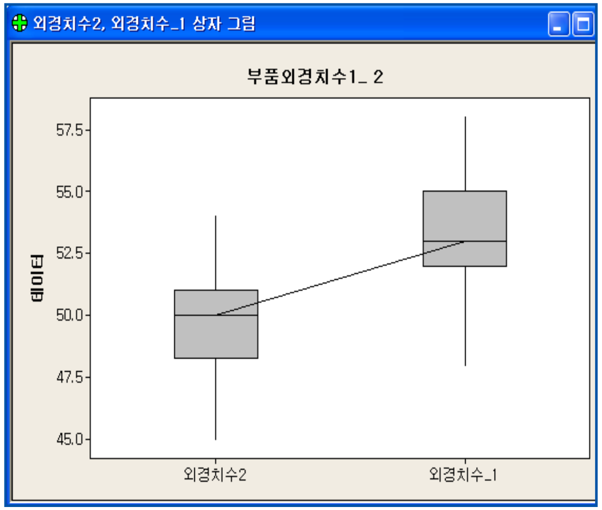
```
줄-잎 그림: 부품외경치수

줄기-잎 그림: 부품외경치수  N  = 100
잎 단위 = 0.10


  1    45  0
  5    46  0000
 12    47  0000000
 25    48  0000000000000
 40    49  000000000000000
(21)   50  000000000000000000000
 39    51  000000000000000000000
 18    52  000000000
  9    53  0000000
  2    54  00
```

1) 자료의 시각적 정리

❑ 상자 그림 (Box Plot)

- 제 일사분위수 (Q1)와 제 삼 사분위수 (Q3)를 네모상자(사분위수)로 연결하고 중앙값을 상자 안에 표시하여 분포의 형태 파악
- 자료분포의 대칭성, 자료의 중심위치, 산포의 정도, 극단점, 이상치 등 분포파악에 효과적으로 이용되는 통계그림
- 여러 집단의 비교에 많이 이용

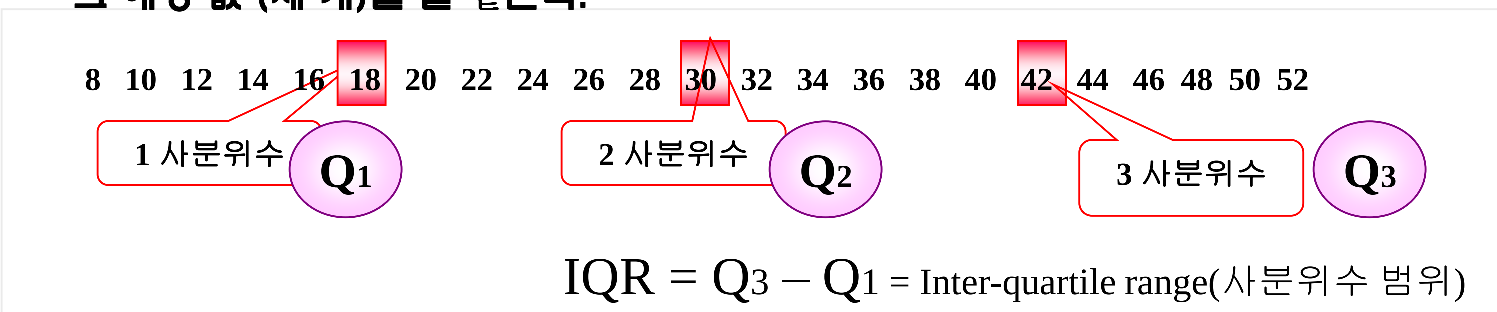
[상자 그림(B)]

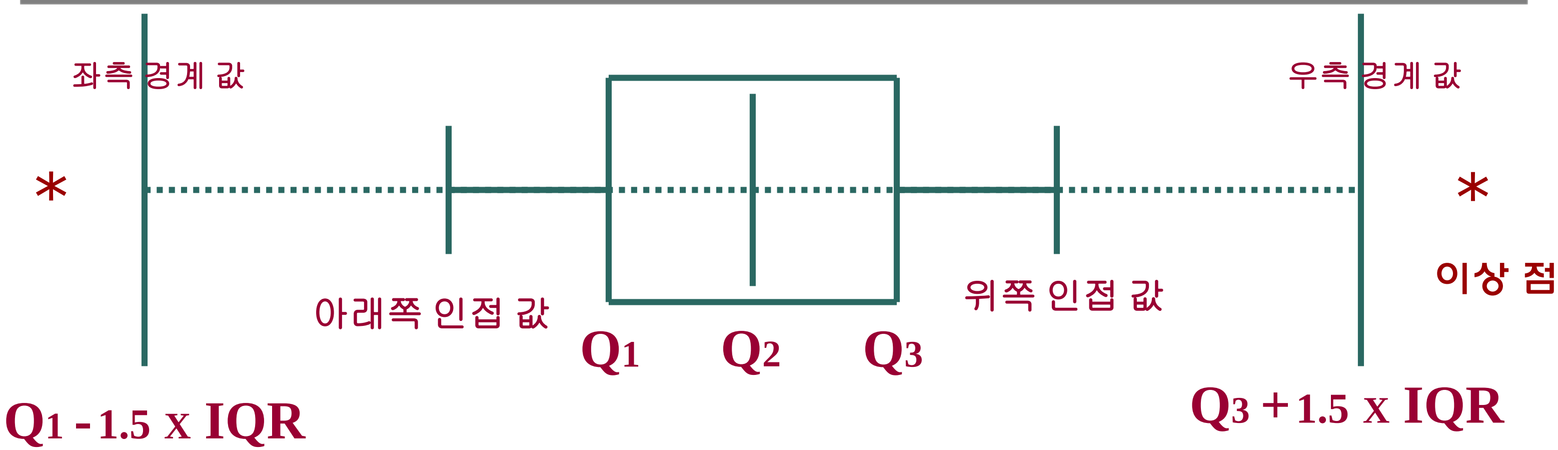


1) 자료의 시각적 정리

❑ 상자 그림 (Box Plot)

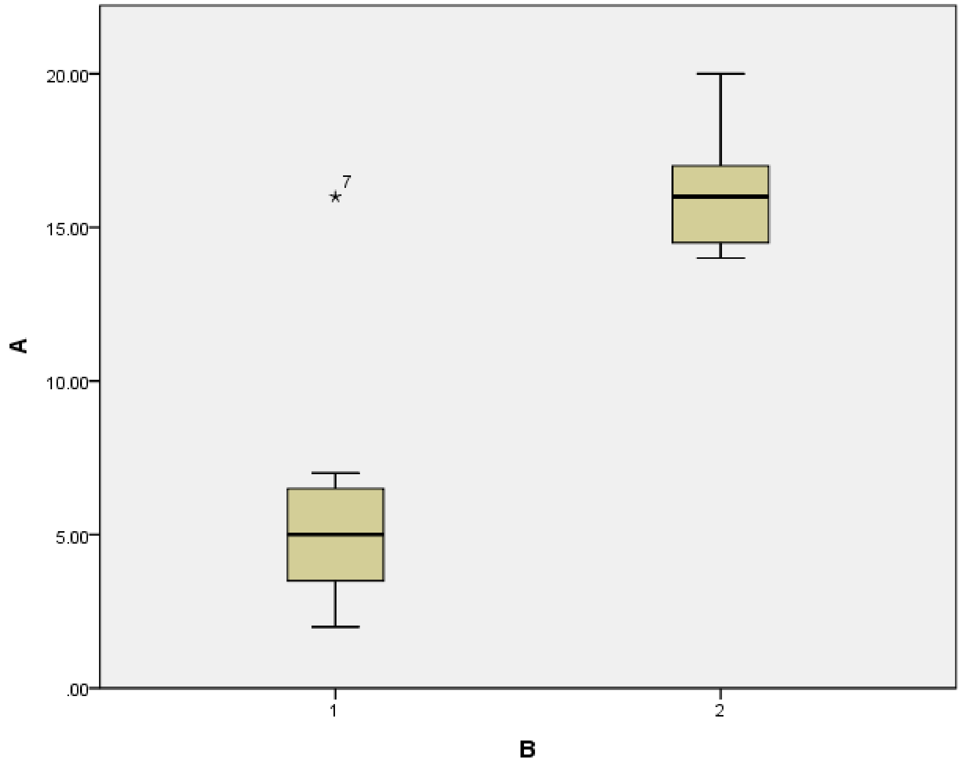
➢ **사분위수란?**
전체 data를 작은 것부터 큰 것으로 순서대로 나열을 하고 이것을 4등분 했을 때 그 해당 값 (세 개)을 일 컬는다.





1) 자료의 시각적 정리

- 상자 그림 (Box Plot)



1) 자료의 시각적 정리

❐ Histogram

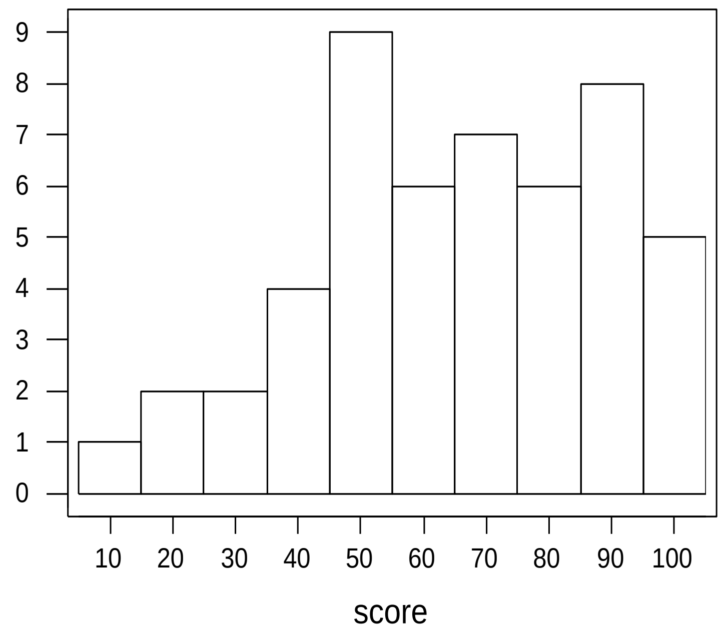
❐데이터가 산포를 가지고 있을 때 어떠한 분포를 하고 있는가를 알아보기 쉽게 발생 빈도수를 그래프로 나타낸 그림이다

❐ 히스토그램은 데이터만으로 알아보기 어려웠던 전체 모습을 간단하게 알 수 있고 데이터의 평균이나 산포의 모습 및 크기를 알 수 있다.

■ 평가점수

65	73	65	36	81
60	43	21	83	64
12	91	60	24	54
69	89	96	86	85
95	85	51	81	47
62	85	46	49	76
44	72	33	46	49
74	78	48	62	97
31	96	97	88	61
54	89	77	72	35

■ Graph> Histogram

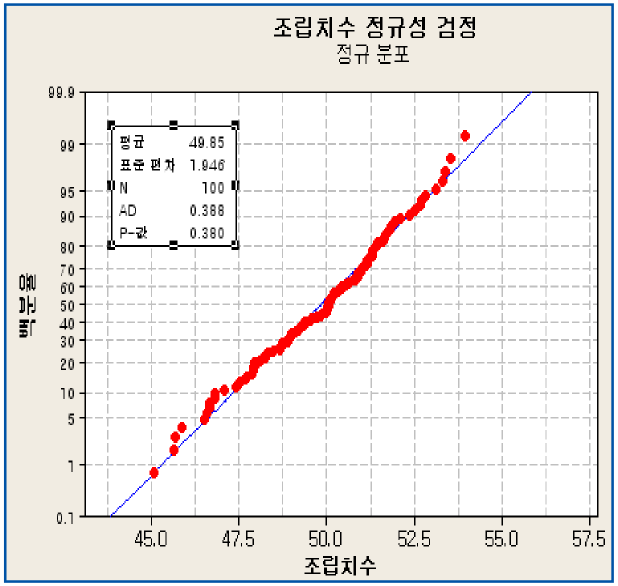


1) 자료의 시각적 정리

❑ 정규 확률도 (Normal Probability Plot)

- 자료가 정규분포를 따르는지 판단하는 그림
- 백분위수 – 백분위수 그림 (Q–Q plot) 방법을 사용
- 정규분포일 경우 직선의 형태. 그 이외의 분포는 구부러진 형태

[정규성 검정(N)]

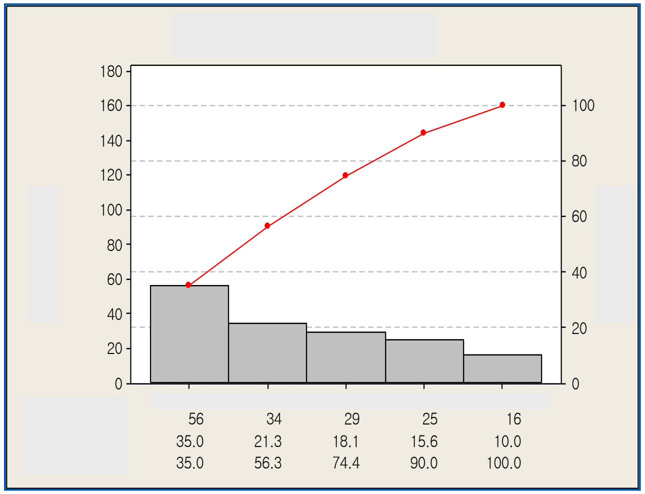


1) 자료의 시각적 정리

❑ 파레토 도표 (Pareto Chart)

- 불량, 고장 등의 발생건수를 항목별로 나눈 후 크기 순서대로 막대그림으로 표시
- 계수형 자료일 때 각 범주에 대한 빈도를 막대의 높이로 나타낸 그림
- 불량품을 발생시키는 원인에 대한 영향정도를 대략적으로 파악할 수 있는 도구

[파레토 도표]



2) 중심위치 측도

❑ 중심위치의 측도

- 평균(Average, Mean) :관측 값들의 합을 관측 값의 총 개수로 나눈 것
- 중앙값(Median) : Data를 크기 순으로 배열했을 때 한 가운데 위치하는 값
- 최빈값(Mode) : Data중 가장 빈도가 많은 값

❑ 중심위치 측도의 특징 비교

- 모집단의 추정치로서의 표준오차 : 평균이 표준오차가 가장 적은 안정성 있는 대표치
- 통계처리의 다양성/계속성 : 대표치 기능이상의 다른 정보를 얻고자 하는 경우 평균계산 필수
- 계산의 간편성 : 최빈값은 분포상에서 즉각적으로 계산
- 자료의 특성 : 좌우대칭이 아닌 극단적인 산포를 이루는 자료는 중앙값이 가장 합당
- 측정수준 : 명목변수는 최빈치, 서열변수는 중앙치, 등간변수와 비율변수는 평균 사용
- 분포상의 비교 : 자료분포가 정규분포인 경우 평균, 중앙값, 최빈값이 일치

2) 중심위치 척도

❒ 평균(Mean ; $\bar{\mu}$,X)

- 관측값을 모두 합한 후에 관측수의 총 수로 나눈 것
- 관측된 데이타의 중심을 측정하는 대표적인 통계량
- 극한값(Outlier)의 영향을 많이 받음

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

중앙치(Median)

측정된 값들을 크기순서대로 정렬했을 때 중앙에 위치하는 값(측정수가 짝수이면 중앙 두개값의 평균)

장점: 극단적인 값에 대해 왜곡되지 않음
단점: 수학적 특성이 결여됨

최빈치(Mode)

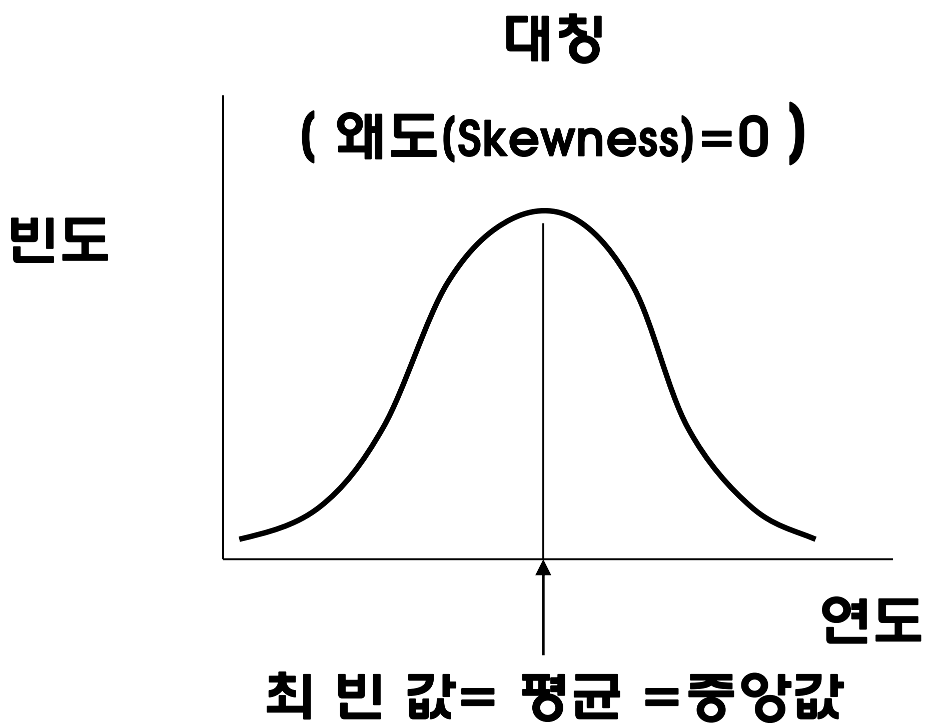
측정된 값에서 가장 빈도가 큰 값

장점: 가장 빈도가 큰 값을 보여줌
단점: 1) 수학적 특성이 결여됨,
2) 경우에 따라 최빈값이 없을 수 있음

2) 중심위치 측도의 선택

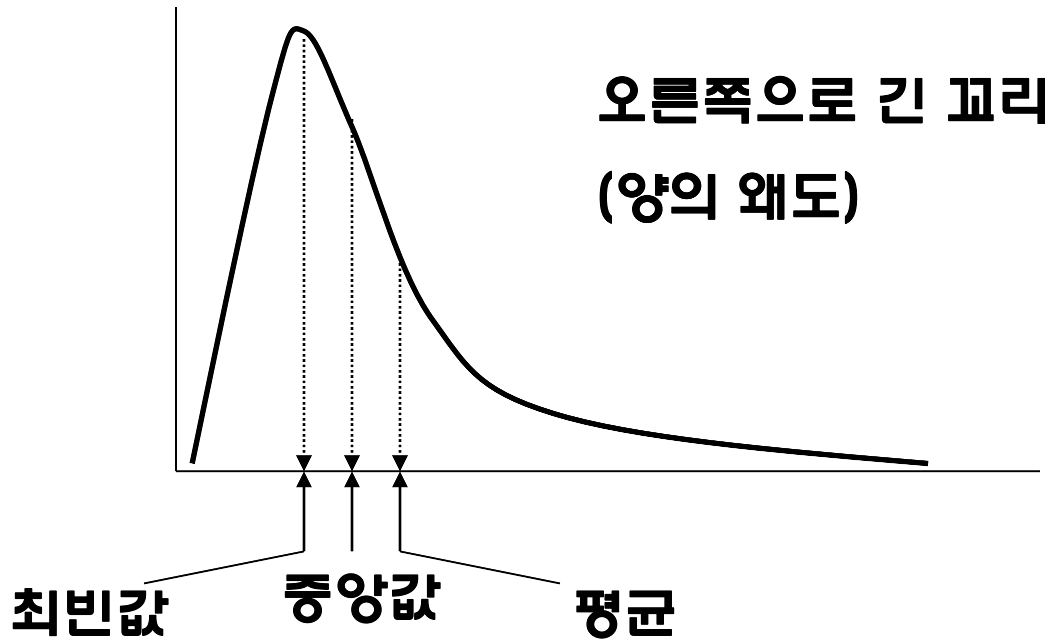
- 대칭분포

$$\text{왜도} = \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3}{n-1}$$



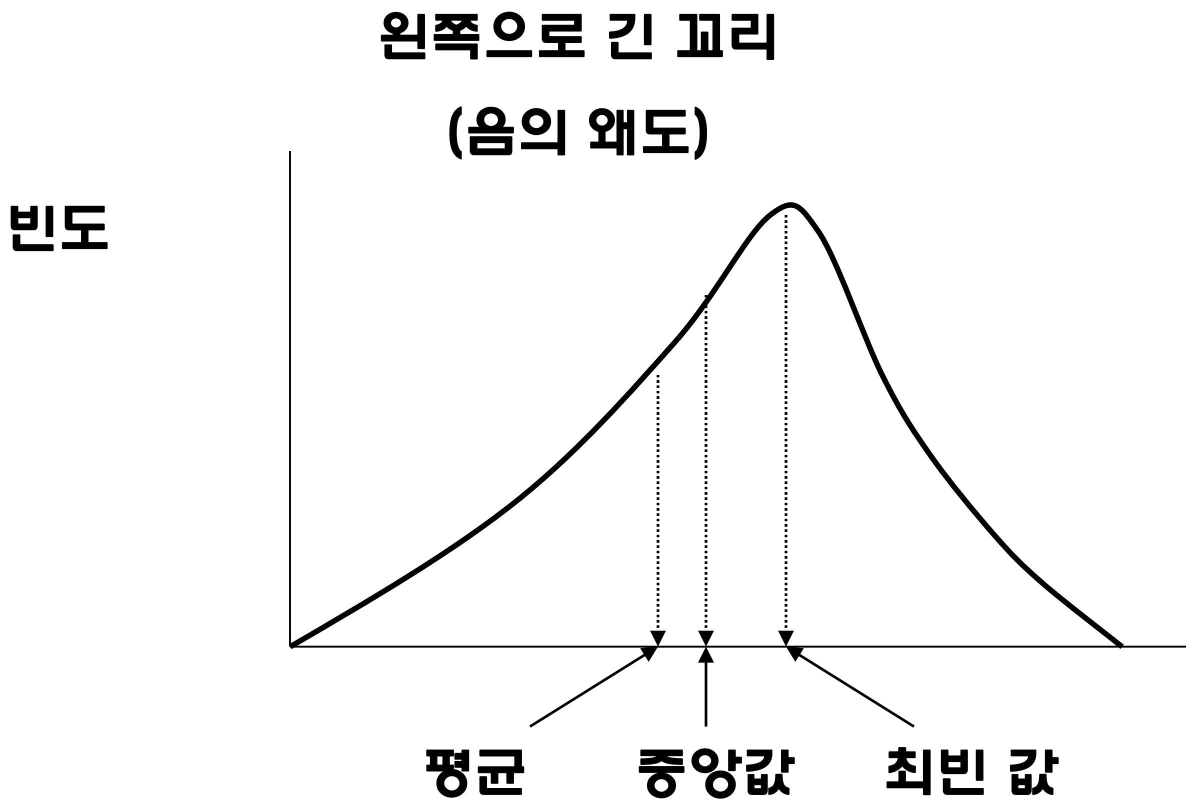
2) 중심위치 측도의 선택

- 오른쪽으로 긴 분포



2) 중심위치 축도의 선택

- **왼쪽으로 긴 분포**



3) 산포의 측도

- 산포(자료들이 중심으로부터 퍼져있는 정도)의 측도
  - 사분위 범위(Inter−Quartile Range) : 3사분위수(Q3) – 1사분위수(Q1)
  - 분산(Variance) : 편차 제곱의 평균
  - 표준편차(Standard Deviation) : 분산의 제곱근
  - 변이계수 : 표준편차를 비교할 때 절대수치보다 상대수치가 필요

$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$	$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$	$CV = \frac{\sigma}{\mu}$
[표준편차]	[분산]	[변이계수]
$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$	$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$	$CV = \frac{\sigma}{\mu}$
[표본표준편차]	[표본분산]	[표본변이계수]

3) 산포의 측도

- 수치적 해석
  - 산포도(퍼짐)
    - 평균 이용 – 분산(variance), 표준편차(standard deviation)
    - 순서대로 나열 – 범위(range), 사분위수범위(IQR)

- 자료에 대한 특성을 언급하려면 ?
  - 대표값과 산포도를 같이 기술해야 함

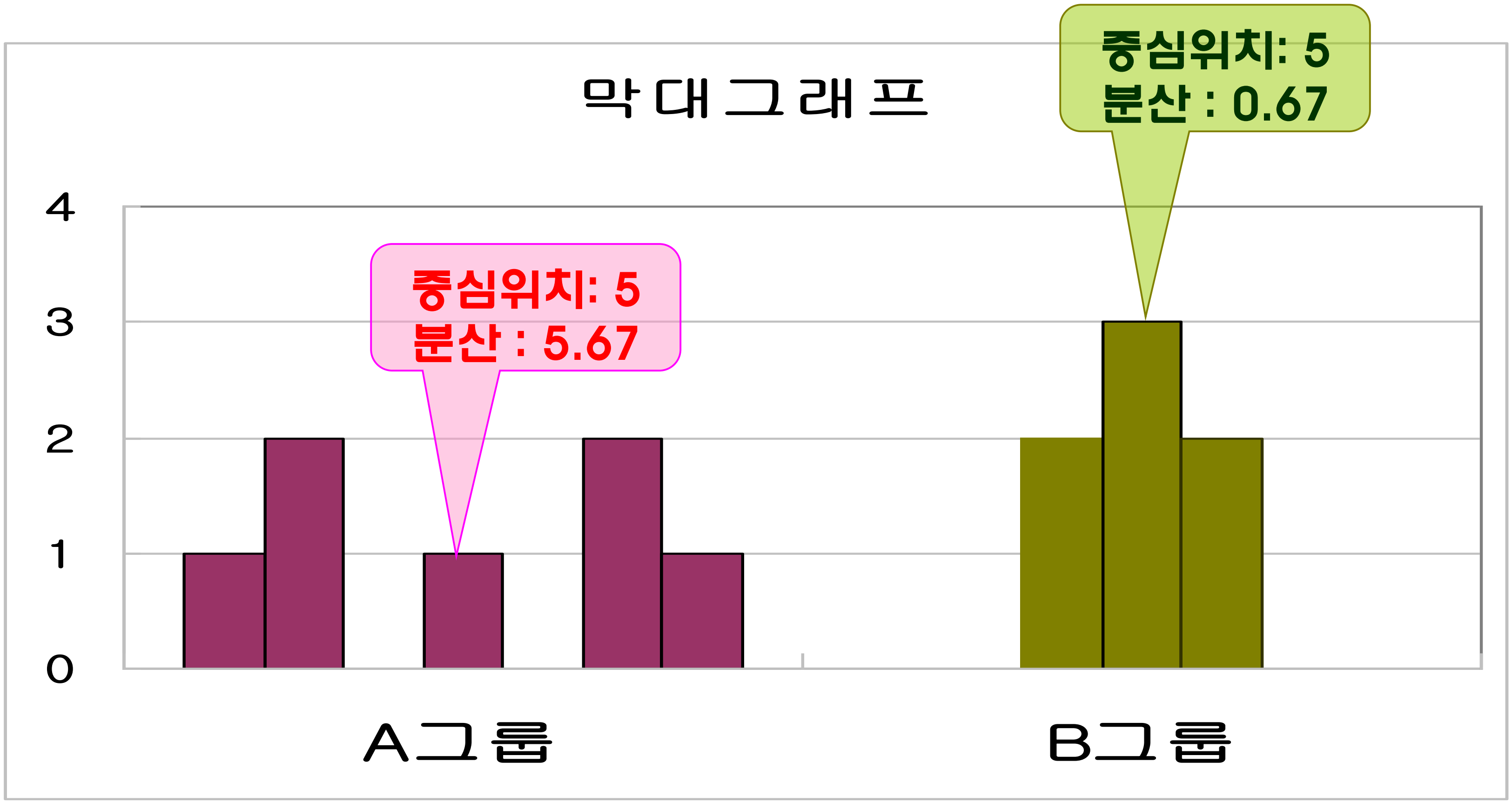
분산은 차이 큼

Group	Data	mean	variance	std.dev
A	2, 3, 3 5, 7, 7, 8	5	5.67	2.38
B	4, 4, 5 5, 5, 6, 6	5	0.67	0.82

평균 동일함

3) 산포의 측도

- 도표적 해석



4) 첨도

- 첨도(Kurtosis) : 자료들의 분포 형태가 중심위치에서 어느 정도 뾰족한가를 나타내는 척도

  - 정규분포보다 뾰족한 봉을 갖는 경우 : 양 (+)의 값
  - 정규분포보다 납작한 봉을 갖는 경우 : 음 (−)의 값

$$첨도 = \frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4}{n-1} - 3$$

5) 산포를 나타내는 척도

❐ 범위(Range ;R)

- 관측된 데이터중 최대값과 최소값과의 차이
- 범위 = 최대값 − 최소값

범위 = 최대값 − 최소값

❐ 분산(Variance; σ^2, S^2)

- 평균과 각 개별 데이터의 차이에 대한 제곱합의 평균
- 데이타의 흩어진 정도를 표현하는 통계량

$$S^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

❐ 표준편차(Standard deviation ; σ, S)

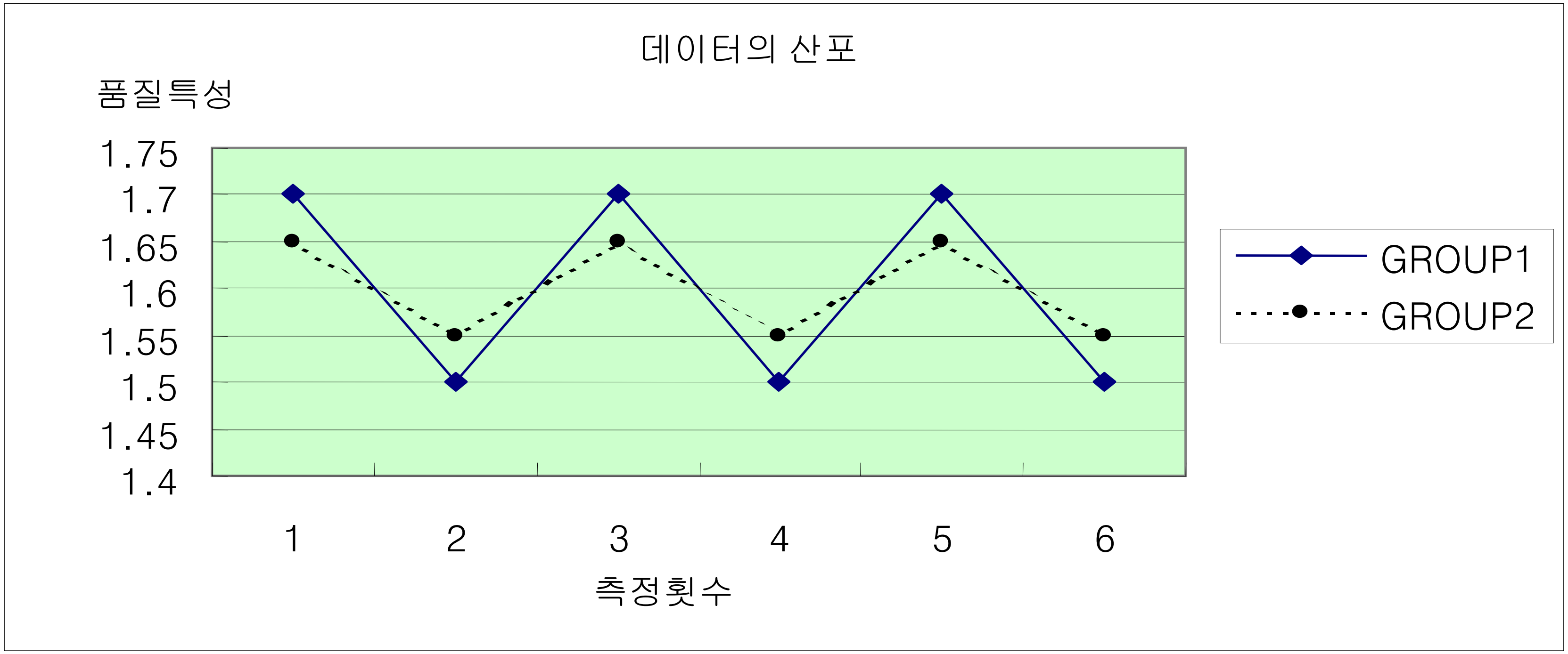
- 분산의 제곱근
- 데이타의 흩어진 정도를 표현하는 보편적인 통계량

$$S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

$$S = \bar{R}/d_2$$

제품의 품질특성과 산포

데이터번호	1	2	3	4	5	6	평균	표준편차
GROUP1	1.7	1.5	1.7	1.5	1.7	1.5	1.6	0.11
GROUP2	1.65	1.55	1.55	1.65	1.55	1.65	1.6	0.05



5) 탐색적 자료분석

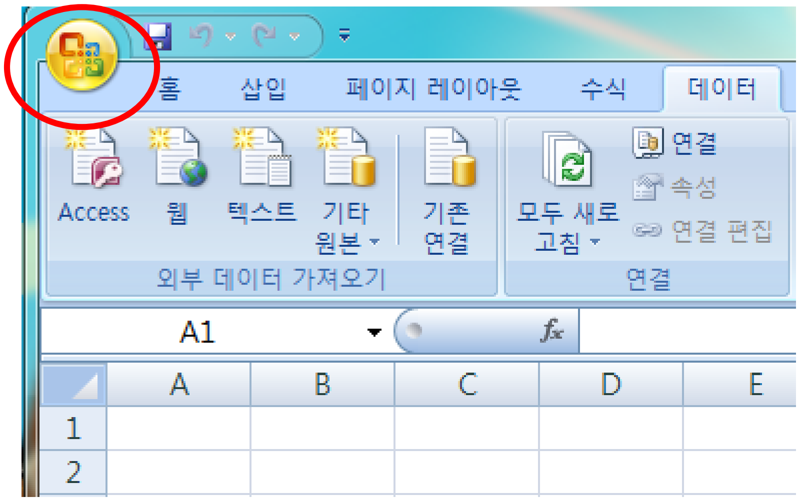
- EDA : Exploratory Data Analysis
  - 각종 그림을 그려본다.
    - ✓ 점 그림, 히스토그램, 상자그림, 산 점 도
  - 자료의 대표 값을 구한다.
    - ✓ 중심 : 평균, 중앙값
    - ✓ 산포 : 분산, 표준편차, 범위, 사분위수 범위
    - ✓ 기타 : 자료의 개수, 최대값, 최소값, 제1사분위수, 제3사분위수

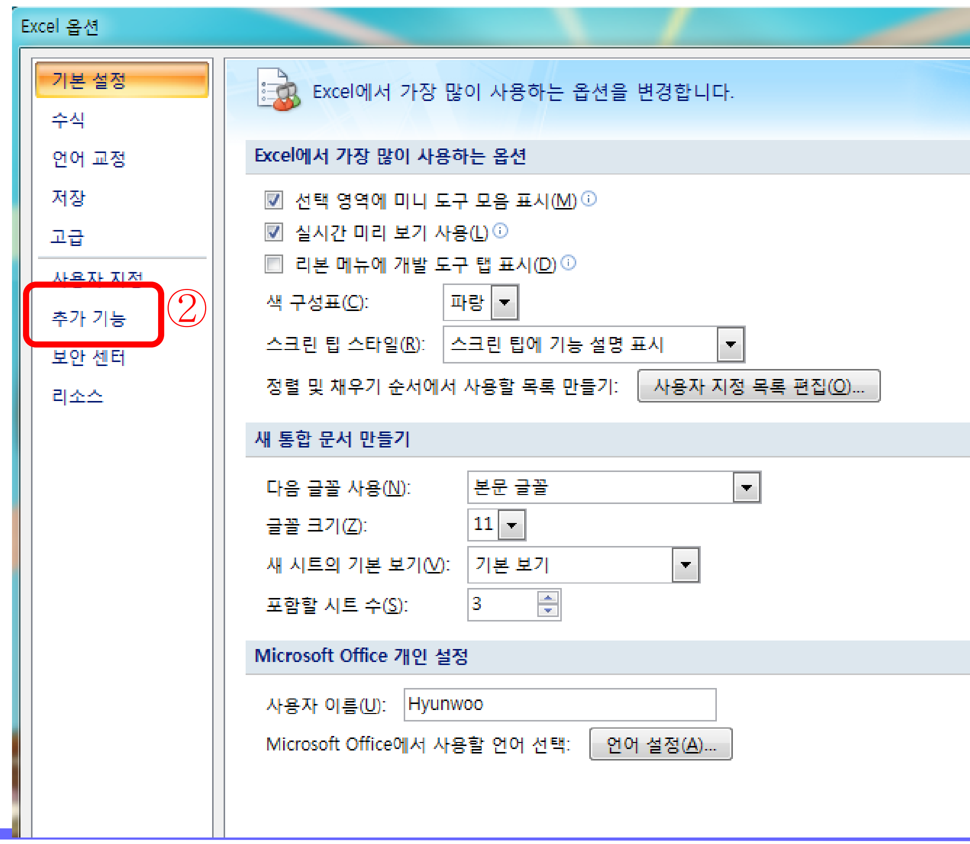
6) 엑셀에서의 통계분석 기능

❑ 데이터 분석 기능 옵션 설정방법

① 엑셀 2007의 옵션설정

2007버전의 형태이며
2003버젼에서는 도구-> 분석도구로
2010, 2013버젼에서는 홈 메뉴에 옵션 항목선택





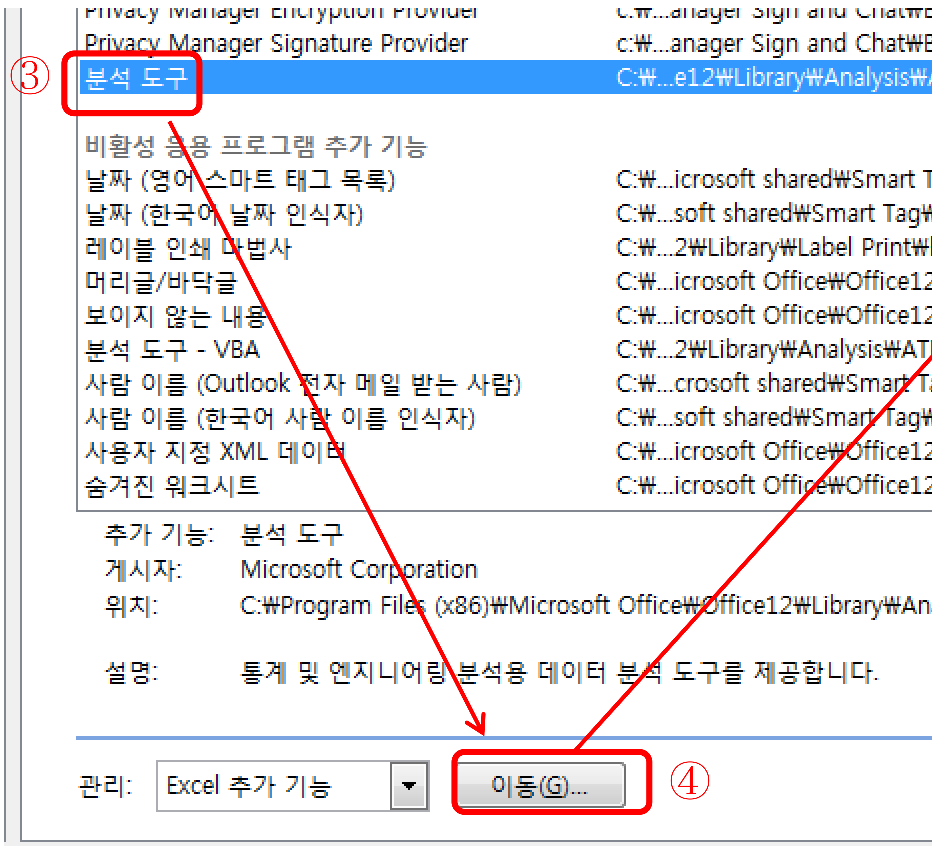
② 추가기능 클릭

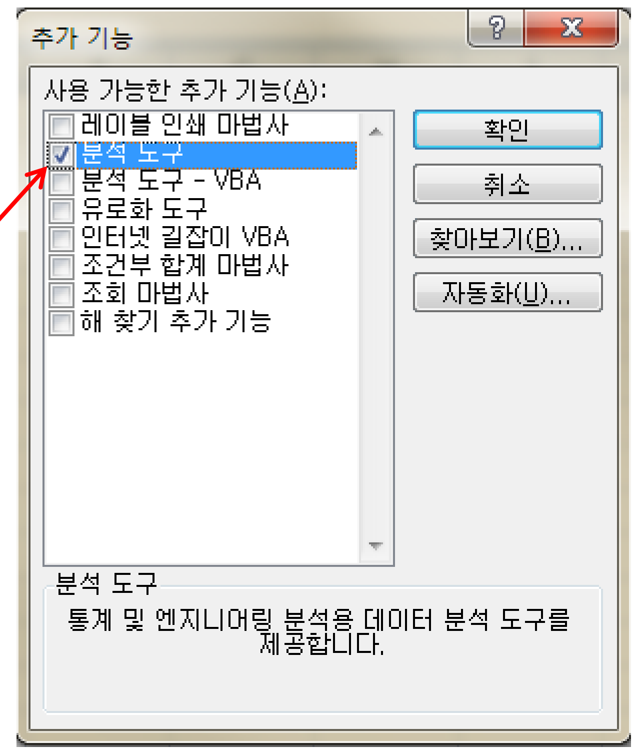
6) 엑셀에서의 통계분석 기능

❑ 데이터 분석 기능 옵션 설정방법

③ 분석도구 클릭 후 이동버튼 클릭

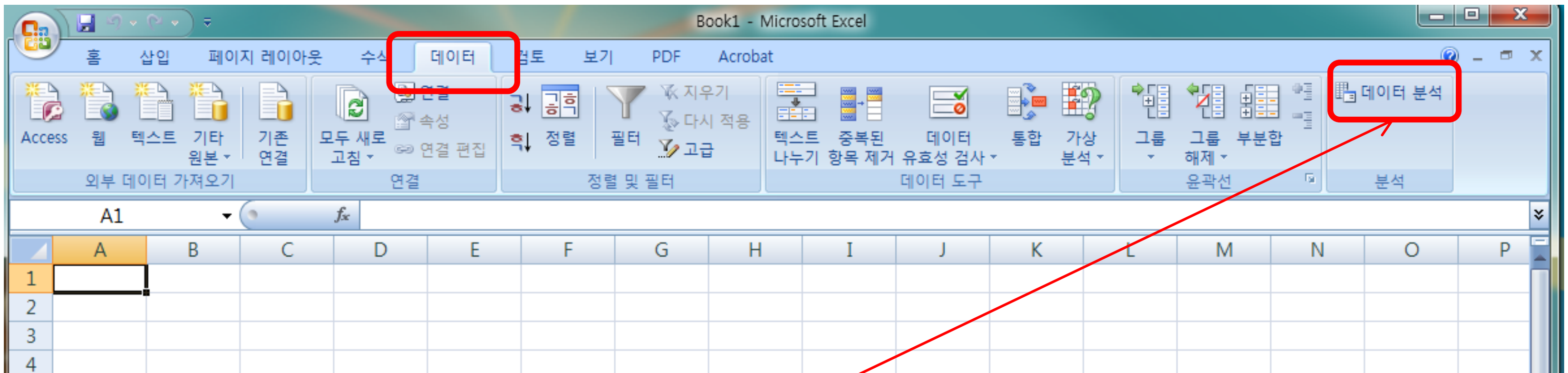
⑤ 분석도구 체크후 확인





6) 엑셀에서의 통계분석 기능

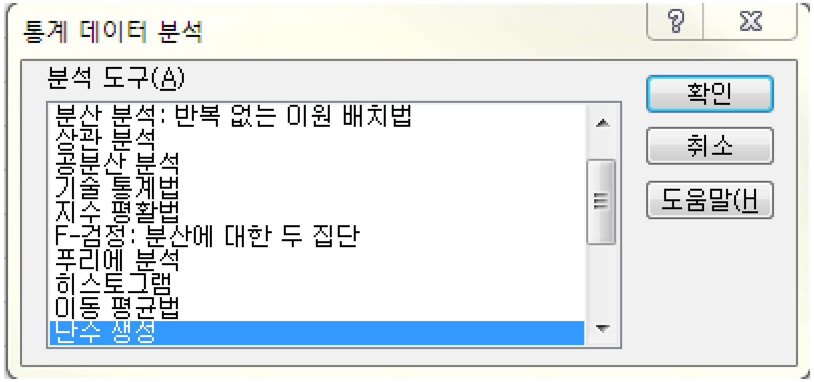
❑ 데이터 분석 기능 옵션 설정방법



데이터 메뉴의 우측 상단에 데이터분석 메뉴가 나타나면 설정은 성공

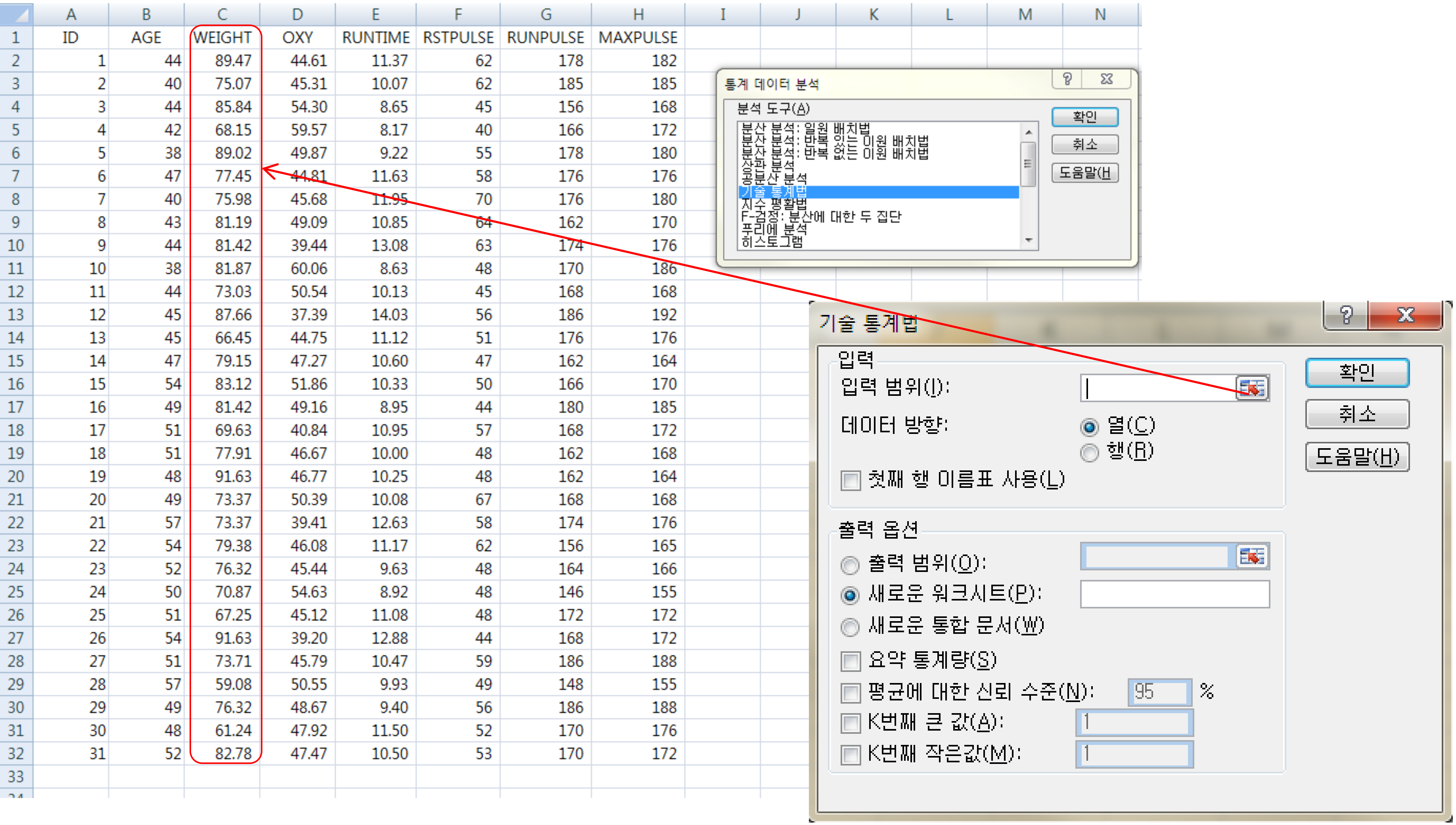
❑ 데이터 분석 기능 옵션 설정방법

데이터 분석 메뉴를 클릭할 경우 아래와 같이 통계데이터 분석창이 뜬다
원하는 분석을 수행



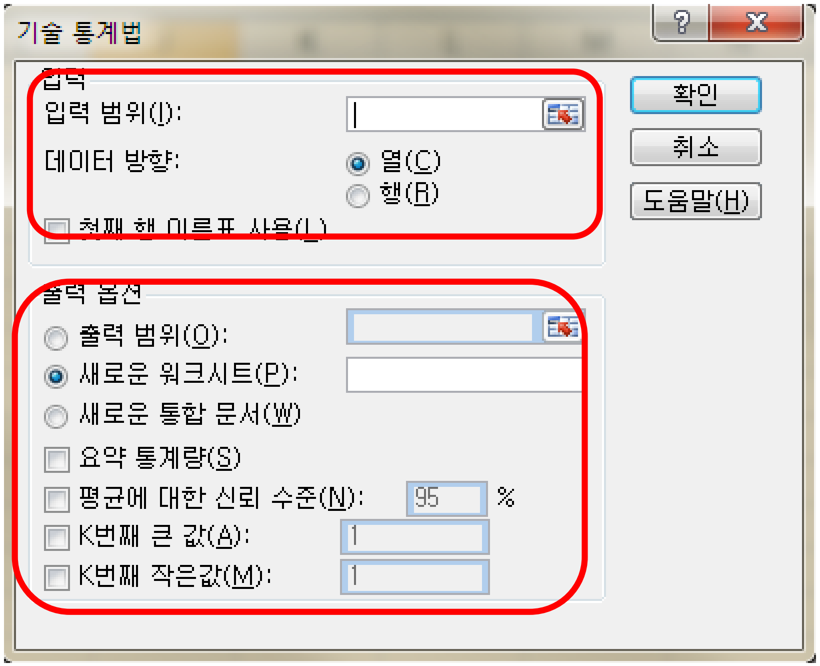
❑ 기술통계법의 사례

입력범위를 클릭 후 분석 하고자 하는 데이터를 긁어서 분석 데이터 선택



ID	AGE	WEIGHT	OXY	RUNTIME	RSTPULSE	RUNPULSE	MAXPULSE
1	44	89.47	44.61	11.37	62	178	182
2	40	75.07	45.31	10.07	62	185	185
3	44	85.84	54.30	8.65	45	156	168
4	42	68.15	59.57	8.17	40	166	172
5	38	89.02	49.87	9.22	55	178	180
6	47	77.45	44.81	11.63	58	176	176
7	40	75.98	45.68	11.95	70	176	180
8	43	81.19	49.09	10.85	64	162	170
9	44	81.42	39.44	13.08	63	174	176
10	38	81.87	60.06	8.63	48	170	186
11	44	73.03	50.54	10.13	45	168	168
12	45	87.66	37.39	14.03	56	186	192
13	45	66.45	44.75	11.12	51	176	176
14	47	79.15	47.27	10.60	47	162	164
15	54	83.12	51.86	10.33	50	166	170
16	49	81.42	49.16	8.95	44	180	185
17	51	69.63	40.84	10.95	57	168	172
18	51	77.91	46.67	10.00	48	162	168
19	48	91.63	46.77	10.25	48	162	164
20	49	73.37	50.39	10.08	67	168	168
21	57	73.37	39.41	12.63	58	174	176
22	54	79.38	46.08	11.17	62	156	165
23	52	76.32	45.44	9.63	48	164	166
24	50	70.87	54.63	8.92	48	146	155
25	51	67.25	45.12	11.08	48	172	172
26	54	91.63	39.20	12.88	44	168	172
27	51	73.71	45.79	10.47	59	186	188
28	57	59.08	50.55	9.93	49	148	155
29	49	76.32	48.67	9.40	56	186	188
30	48	61.24	47.92	11.50	52	170	176
31	52	82.78	47.47	10.50	53	170	172

엑셀에서의 통계분석



입력부분 : 분석하고자 하는 데이터를 지정
첫 번째 행 이름표 사용 : 분석 변수의 변수명이 데이터의 첫번째 행에 있을 경우에 사용

출력부분 : 분석결과을 어디에 제공할 것인가를 정리하고, 분석 통계량에 대한 내용을 지정
출력지정 : 선택한 셀 부터 출력 결과 제공
새로운 워크쉬트 : 새로운 워크쉬트에 결과 제공
새로운 통합문서 : 새로운 엑셀파일에 결과 제공
요약통계량 : 필수적으로 선택

대부분의 통계분석에서 동일한 형태로 사용되고 있음

확률 변수

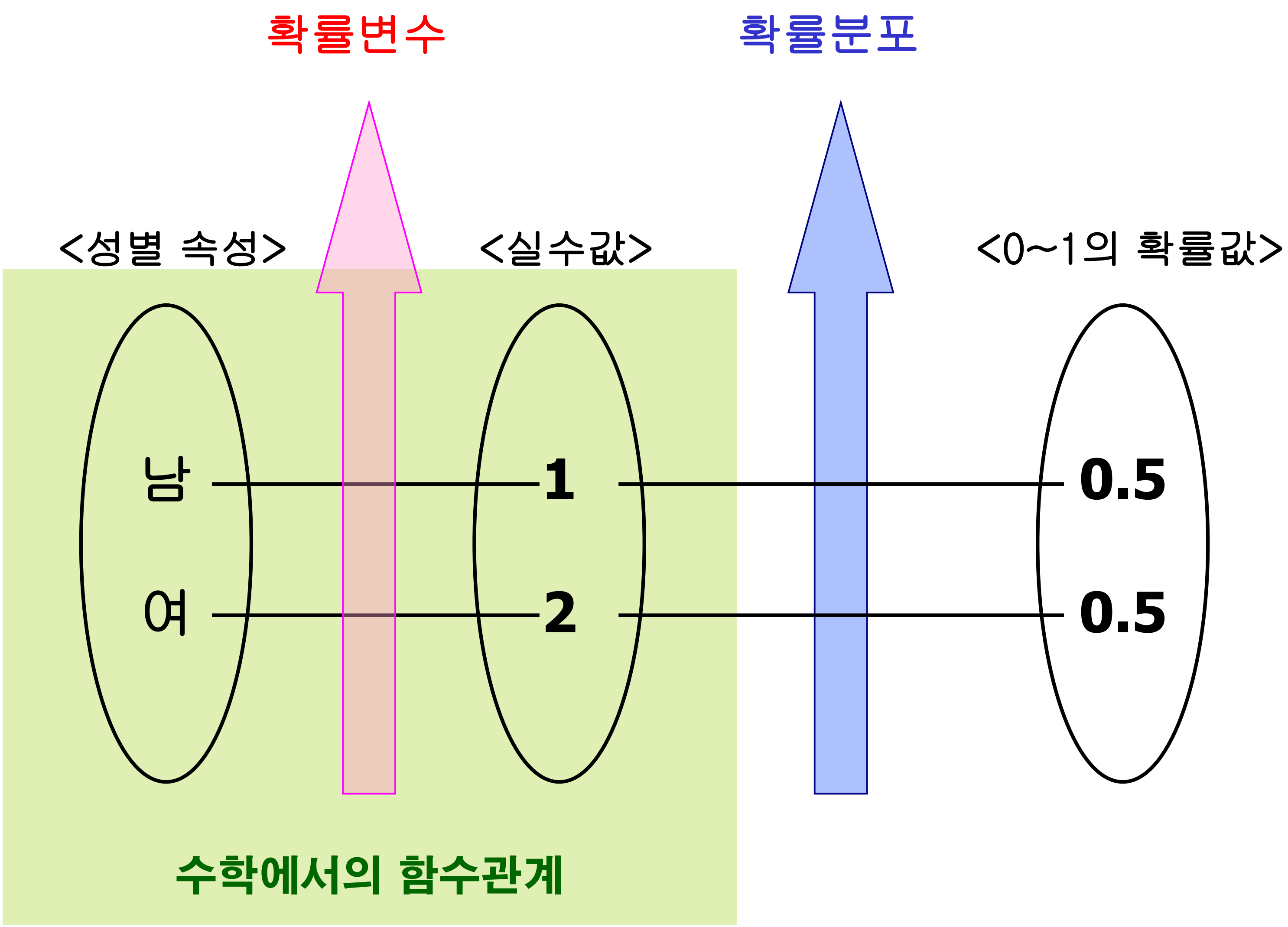
- **통계적 실험(Statistical Experiment)**
  - 비슷한 사건의 반복으로 여러 가지 가능한 결과가 있을 수 있지만, 정확히 무슨 결과가 발생할지는 모르는 현상은 통계학의 연구 및 응용 대상이 된다.
  - 이런 현상에 대한 통계학적 연구를 통계학적 실험이라고 한다.
  - 표본공간 : 불확실성을 구체적으로 표현하는 것으로서 관찰 가능한 모든 가능한 집합. 이산형/연속형 표본공간(자료의 구분과 동일)

- **확률변수(Random Variable)**
  - 표본공간을 대상으로 직접 문제 해결이 곤란한 경우 표본공간을 수직선 위로 변환
  - 정의: 표본공간에서 정의된 실수치 함수
  - 예: 동전을 3회 던지는 실험
  - 이산형 확률변수, 연속형 확률변수

확률 변수란?

	확률변수	함수
비교	– 분야 : 통계 – 표기 : X, Y, Z – 의미 : 개체속성들을 실수값에 대응시키는 것 – 이산형 확률변수, 연속형 확률변수	– 분야 : 수학 – 표기 : y, f(x) – 의미 : 집합 A의 원소를 집합 B의 원소에 대응시키는 것
확률분포	확률변수의 값(또는 확률함수)과 그에 대응하는 확률(또는 구간확률)을 대응시키는 것	

확률 변수란?



확률분포의 종류

- 확률밀도함수 (pdf: probability density function)
  - 확률변수 X의 분포를 나타내는 함수
  - 이산분포함수로는 이항분포, 포아송분포 등이 있고,
  - 연속분포함수로는 정규분포, 카이제곱분포 등이 있다.
  - 이산분포 pdf는 보통 p(x)로, 연속분포 pdf는 보통 f(x)로 나타낸다.

- 이항분포(Binomial Distribution)
  - 실험을 n번 실시하여 얻은 실험결과 중에 '성공의 회수'를 X라 할 때
  - X가 취할 수 있는 값은 0,1,2,…n으로 이항분포에 따른다.
  - $p(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0,1,2,\ldots\ldots,n$
  - $E(x) = np, V(x) = np(1-p)$

포아송 분포 (Poisson Distribution)

- 단위시간당 발생하는 한 사건 ('전화가 걸려옴', '교통사고 발생', '기계 고장')의 수를 조사
- 단위시간당 '성공의 회수' 가 평균 m 이라 할 때 포아송 확률변수 X= '단위시간당 성공 회수 ' 의 분포는

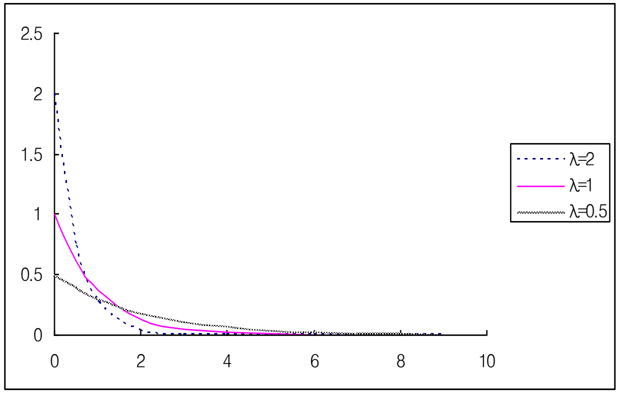
$$p(x) = \frac{e^{-m} m^x}{x!}$$ 평균 : E(X) = m , 분산 : V(X) = m

- 이항분포와 포아송 분포와의 밀접한 관계
  - n 이 무한대에 접근하고 p가 0에 접근하여 평균 성공수 np=m 는 일정한 상수인 경우에는 이항분포는 포아송 분포로 근사하게 구할 수 있음.
- 포아송 분포의 사례
  - 어느 회사 부서 사무실에 오전 9시에서 10시 사이에 걸려오는 전화의 수
  - 어느 교차로에서 발생하는 1일 교통사고의 수
  - 옷감의 단위 길이당 발생하는 결점 수

지수 분포 (Exponential Distribution)

신뢰성에서 가장 많이 사용되는 분포

시간이 지남에 따라 고장률이 일정 한 어떤 제품이 고장이 일어나고 다음 고장이 일어날 때 까지 걸리는 시간



- 확률밀도함수

$$f(t) = \lambda e^{-\lambda t}, t > 0 \quad \lambda\text{: 고장률}$$

- 분포함수(불신도함수) 및 신뢰도함수

$$F(t) = \int_0^t f(t)dt = 1 - e^{-\lambda t}$$

$$R(t) = 1 - F(t) = e^{-\lambda t}$$

- 평균 및 분산

$$\text{MTTF} = \int_0^\infty R(t)dt = \frac{1}{\lambda} = \theta$$

$$Var(T) = \frac{1}{\lambda^2} = \theta^2$$

- 고장률함수

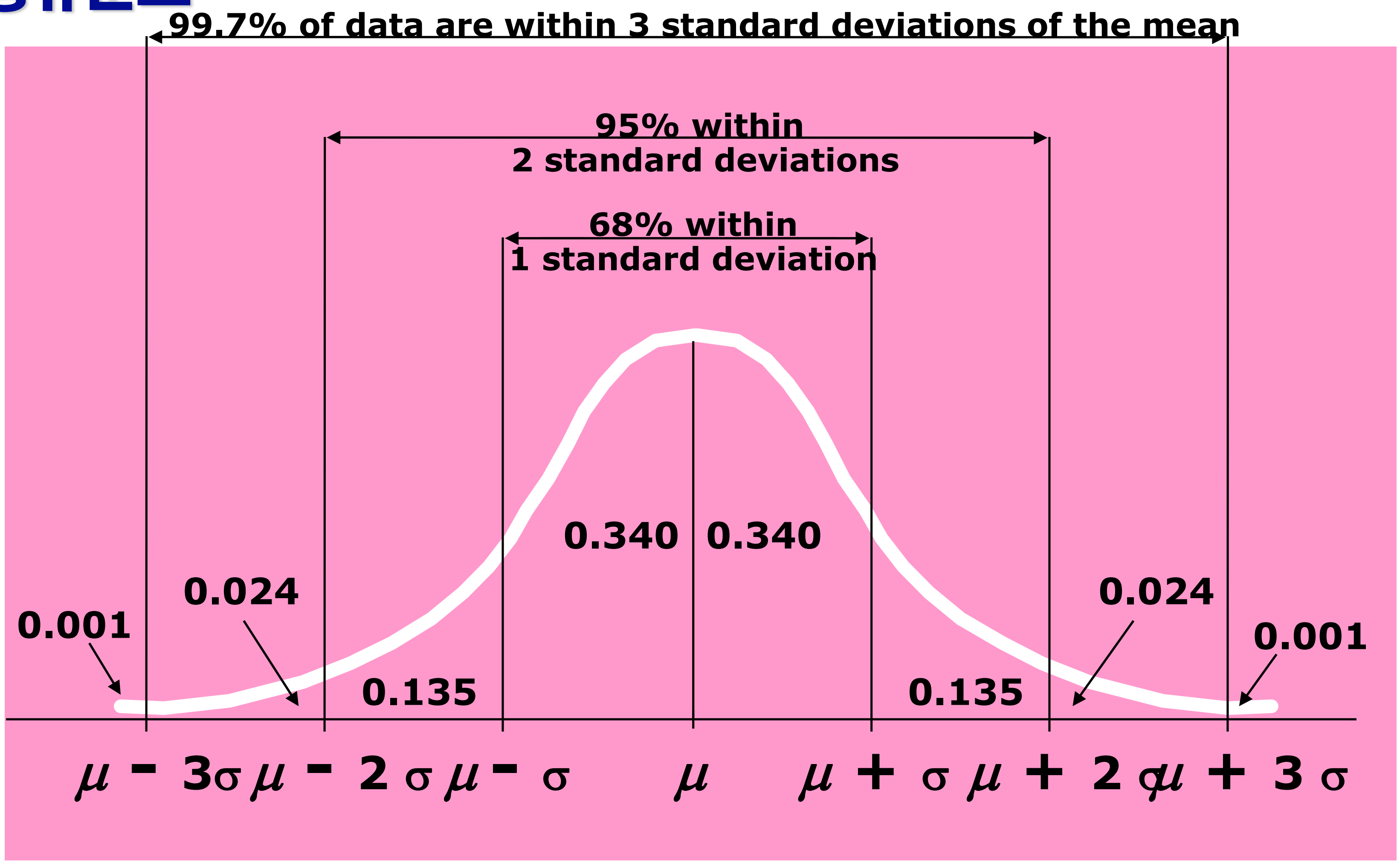
$$\lambda(t) = \frac{f(t)}{R(t)} = \lambda$$

- 지수분포의 고장률은 시간과는 무관하게 상수(λ)
- 평균수명 MTTF(θ)와 고장률 λ는 역수관계

- 백분위수 t_p

$$F(t_p) = 1 - e^{-\lambda t_p} = p, \quad t_p = \frac{1}{\lambda}\{-\ln(1-p)\}$$

정규분포

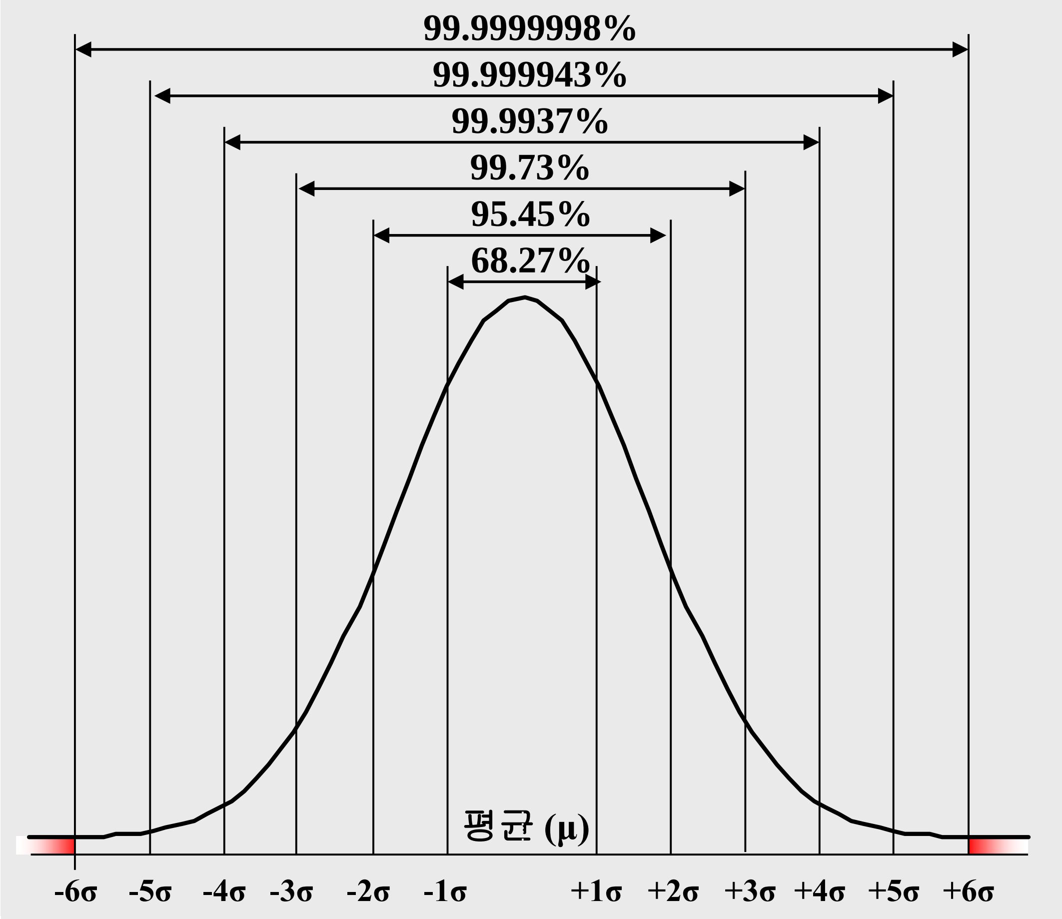

99.7% of data are within 3 standard deviations of the mean
95% within
2 standard deviations
68% within
1 standard deviation
0.340
0.340
0.001
0.024
0.135
0.135
0.024
0.001
μ - 3σ
μ - 2σ
μ - σ
μ
μ + σ
μ + 2σ
μ + 3σ

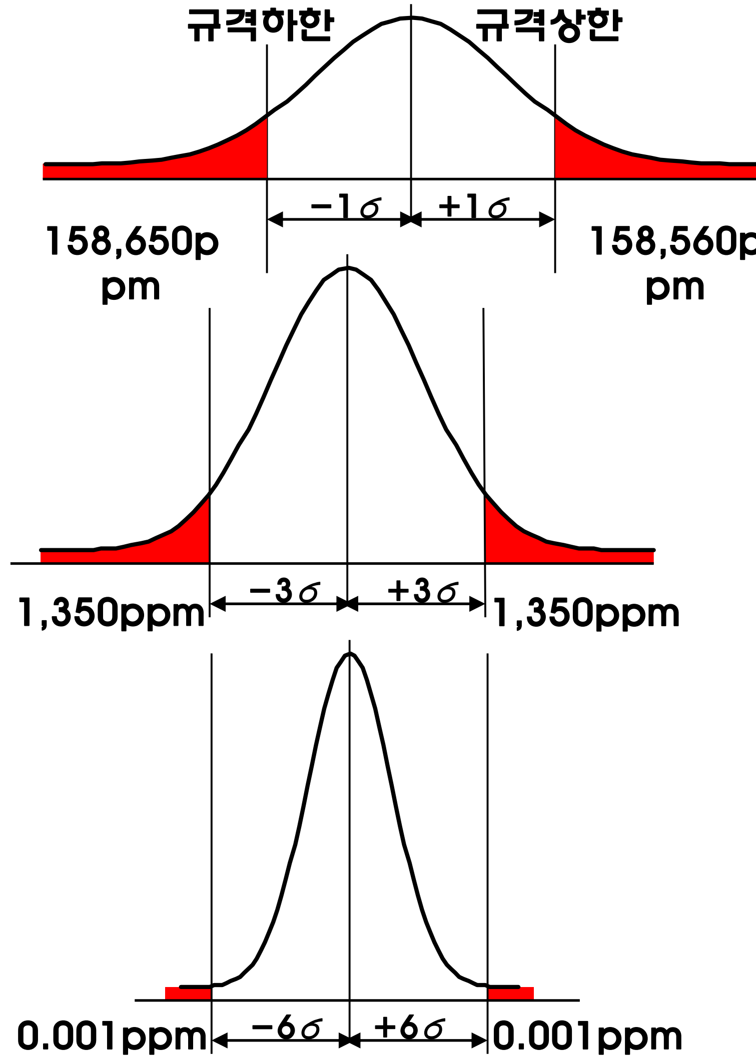
정규 분포

- 무한대의 샘플을 측정하여 얻을 수 있는 이론적인 분포
- 분포의 형태가 종을 엎어 놓은 모양이며,
- 평균값을 중심으로 좌,우 대칭으로
- 평균(μ) 와 분산(σ^2)에 의하여 위치와 산포가 결정된다.
  - ***히스토그램은 표본(sample)을 사용하여 작성된다.***
  - ***표본통계(x,s)는 표본에서 계산된다.***
  - ***히스토그램과 표본통계를 가지고 이 표본을 추출한***
  - ***모집단을 나타내는 곡선을 만들어 낸다.***
- 표본 데이터가 정규 분포를 하고 있으면 정규분포 곡선을 이용하여 정확한 통계적인 분석을 할 수 있다. (추정통계의 배경)

정규 분포 와 6시그마 공정

공정의 평균에서 규격의 경계치까지의 거리가 표준편차(σ)의 6배되는 거리에 있다는 뜻.





정규분포

표준정규 값 $$Z = \frac{X - \mu}{\sigma}$$

- 평균이 0이고 표준편차가 1인 정규분포를 표준정규분포라 한다.
- X가 평균이 μ이고 표준편차가 σ인 정규분포를 따를 때, Z는 평균이 0이고 표준편차가 1인 정규분포를 따른다.

중심극한 정리

❑ 평균에 대한 중심극한 정리

➢ $X_1, \cdots X_n$을 평균이 μ 이고 분산이 σ인 모집단으로 부터 구하여진 표본이라 하면 , $\bar{X}$ 의 분포는 근사적으로 N(μ , σ^2/ n)에 따르고

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$ 은 근사적으로 N (0 ,1)을 따른다.

❑ 모 비율에 대한 중심극한 정리

➢ X가 B (n , p)이고 n이 크면 $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ 는 근사적으로 N(0,1)을 따른다.

여기서, $\hat{p} = \frac{x}{n}$

정규분포의 연습

- 어느 과수원에서 수확되는 사과의 무게는 평균 200g, 표준편차는 25g인 정규분포를 따른다고 한다. 사과 무게를 X라 하면, 이 과수원에서 수확되는 사과 중 무게가 150g 미만인 것은 하품으로, 260g 이상은 상품으로 간주한다고 한다. 이 과수원에서는 일년에 15,000개의 사과를 수확한다고 한다.

(1) 이 과수원에서 예상되는 하품 사과의 수는 얼마인가?
(2) 이 과수원에서 예상되는 상품 사과의 수는 얼마인가?
(3) 사과의 무게가 190g에서 210g 사이에 있는 사과의 수는 얼마인가?

모집단(Population)

- 조사하고자 하는 대상집단 전체
- 전체조사는 많은 시간과 비용소요

✓ 현재까지 생산된 모든 쏘나타 차량의 평균 중량

✓ 우리나라 총 유권자의 정당별 선호도

표본(Sample)

- 조사하기 위하여 뽑은 일부 집단
- 조사대상 모집단의 부분집합

✓ 2000년 4월 생산된 쏘나타 차량중 50대의 평균중량

✓ 전국의 유권자 1,500명을 대상으로 조사한 정당별 선호도

1.5 표본과 표본분포

표본추출(Sampling)

- **사용이유**
  - 모집단 전체를 조사하는 것이 불가능하거나 어려운 경우
  - 표본추출을 통해 모집단에 대한 효율적인 정보수집

- **확률추출법/ 비확률 추출법**
  - 확률추출법
    - ✓ 모집단으로부터 구성원을 추출하는 과정이 무작위하게 이루어지는 방법
  - 비확률 추출법
    - ✓ 개인적인 판단이나 편의에 따라 모집단으로부터 구성원을 추출하는 과정
    - ✓ 표본으로부터 모집단의 결론에 대한 신뢰도가 객관적 척도가 불가능

Sampling

- **표본추출오차/ 비표본 추출오차**
  - 표본추출오차 – 우연오차, 편의
    - ✓ 표본선택방법과 관련된 오차
  - 비 표본 추출법
    - ✓ 잠재적인 응답자들이 동일한 확률로 뽑혔다고 확신할 수 없음.
    - ✓ 측정방법, 과정의 부정확으로 인한 오차
    - ✓ 측정계기의 부정확, 측정기술의 부족 등으로 인한 오차
    - ✓ 표본오차를 추정할 수 없기 때문에 일반화하여 사용할 수 없음.

- **단순랜덤화 추출법(Simple Random Sampling)**
  - 모집단에 포함되어 있는 모든 구성원이 뽑힐 확률을 같게 하여 뽑는 방법
  - 주사위 같은 기구를 사용하거나, 모집단이 클 경우 난수표를 이용
  - 여러 표본추출방법 중에서 가장 기본이 되며, 다른 추출방법에 응용이 많이 됨.

Sampling

- **층화 추출법(Stratified Sampling)**
  - 모집단의 성격에 따라 여러 개의 층으로 분류한 다음 각 층에서 단순 랜덤화 추출법에 의해 추출
  - 층내에서 동질성이 높고 층간에는 이질성이 높을 때 정확도가 더 높음.

- **집락 추출법(Cluster Sampling)**
  - 모집단이 자연적으로나 인위적으로 집락(cluster)을 형성하고 있을 경우
  - 집락 중 몇 개를 랜덤 하게 선택하여 전수를 조사하는 것
  - 모집단이 크고 넓게 퍼져 있을 때 효과적

- **계통 추출법(Systematic Sampling)**
  - 공간적으로 혹은 시간적으로 일정한 간격으로 추출하는 방법
  - 첫번째 표본은 랜덤하게 추출하고 두번째부터는 일정한 시간적/공간적 간격을 두고 추출
  - 경향성이나 주기성이 있는 경우 편의가 클 가능성이 있음.
  - 단순확률추출보다 표본추출작업이 용이하여 비전문가도 쉽게 이용
  - 단순확률추출법에 비해 일반적으로 단위비용 당 얻는 정보의 양이 더 많음.

표본 분포와 표본오차

- 모수와 통계량
  - 모 수(parameter): 모집단의 특성을 나타내는 수치로서 고정된 값 이지만 대부분은 모르기 때문에 가정을 하거나 추정을 한다.
    - ✓ 예) 모평균, 모 분산, 모 비율
  - 통계량(statistic): 표본으로 부터 계산 되는 값으로서 어떤 개체가 표본으로 추출되냐에 따라 값은 변한다.
    - ✓ 예) 표본평균, 표본분산, 표본비율
- 표본 분포란 ?..... 정확한 표현은 통계량의 표본분포는?
- 표준오차란? …… 정확한 표현은 통계량의 표준오차는?

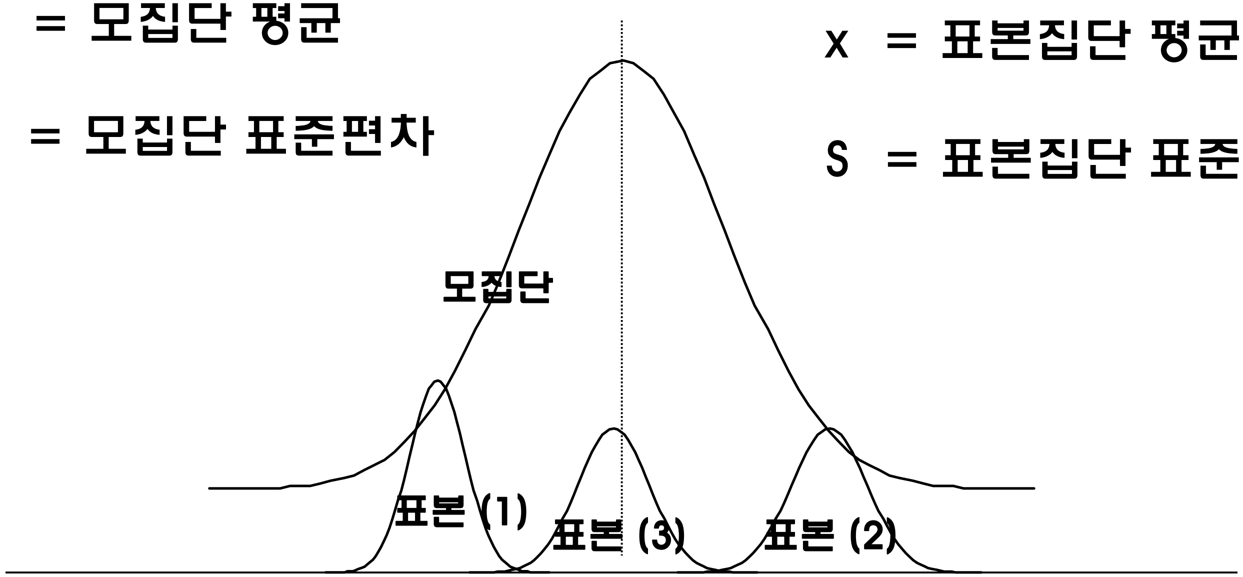
- 모수(Population Parameters)
- 표본통계(Sample Statistics)

μ = 모집단 평균

σ = 모집단 표준편차

$\bar{x}$ = 표본집단 평균

S = 표본집단 표준편차

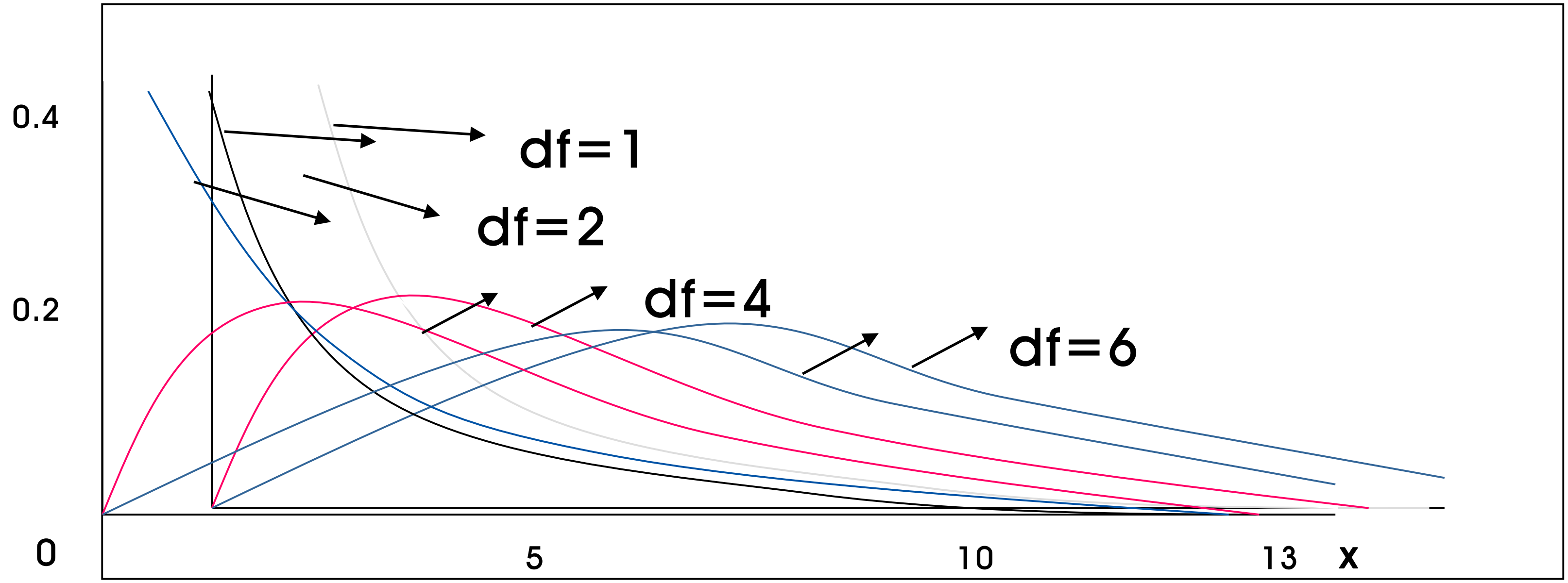


註) 모수(Parameter) – 표본관측에 의하여 구하고자 하는 모집단의 특성 값

여러 가지 표본 분포들

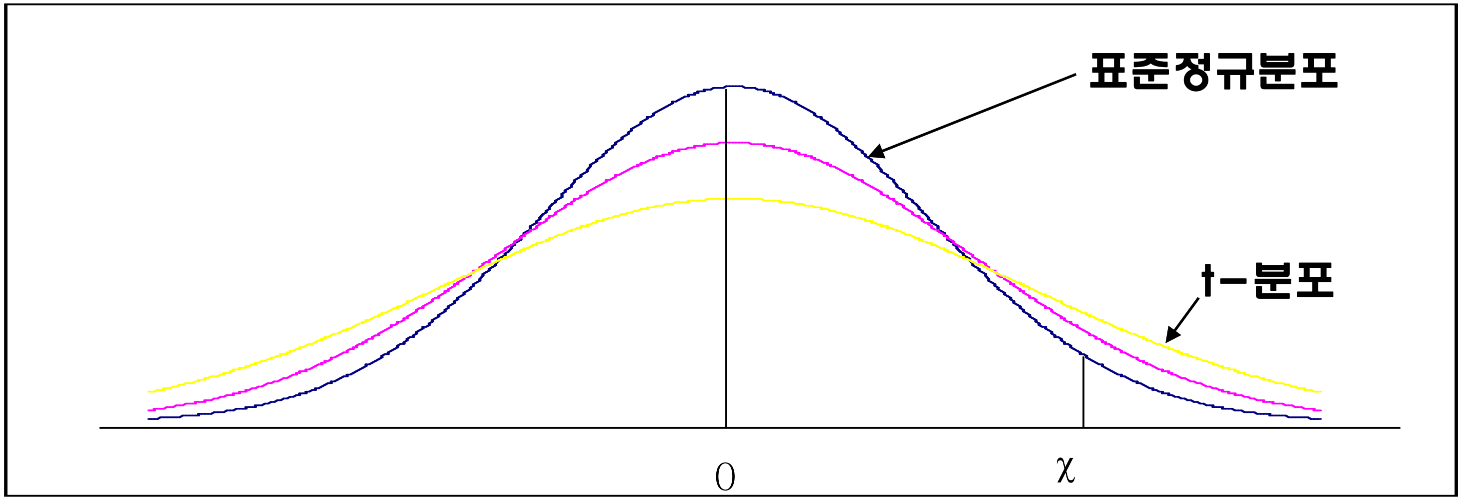
❑ 카이제곱 분포

- 정규분포를 따르는 모집단에서 표본을 추출했을 때, 표본분산의 분포가 카이제곱 분포이다.
- 모 분산에 대한 추론, 범주형 자료의 분석 등에 유용하게 활용
- 비 대칭분포이며 모수인 자유도가 변함에 따라 분포가 달라짐
- 자유도가 많아질수록 정규분포에 근사



여러 가지 표본 분포들

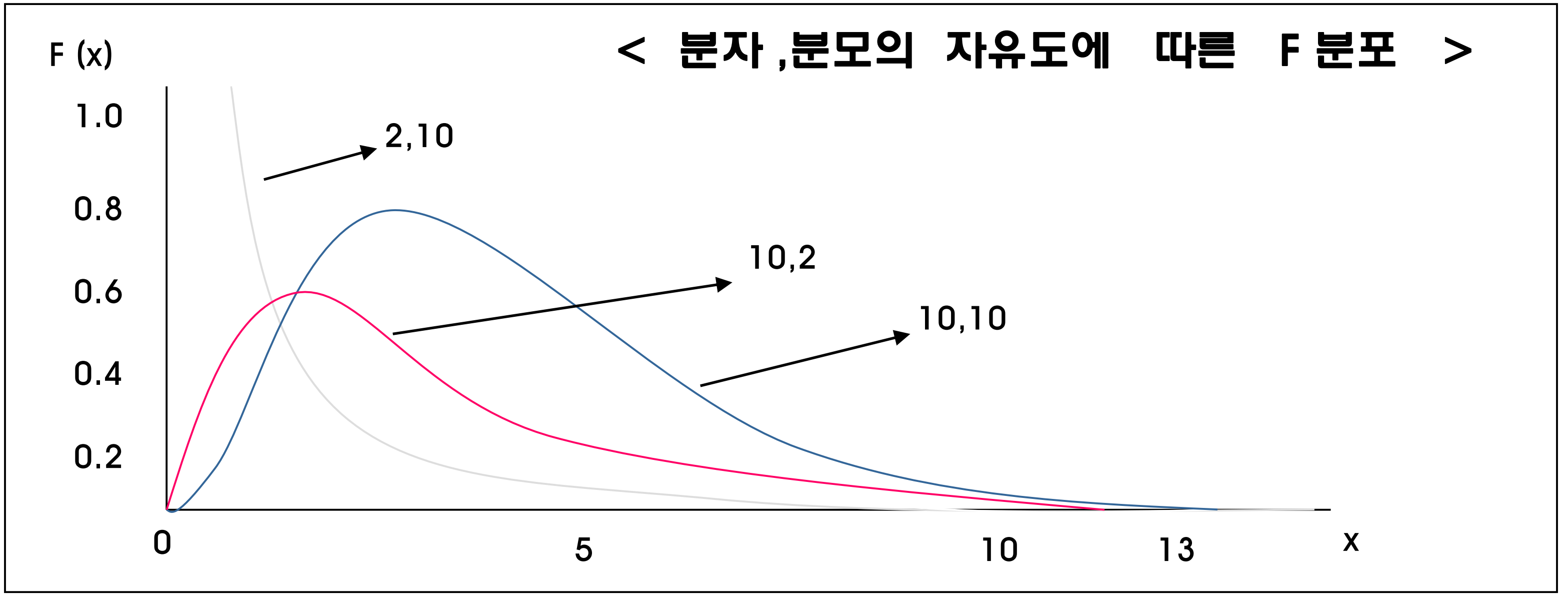
- t 분포
  - 정규분포를 따르는 모집단에서 표본을 추출했을 때, 표본 표준편차를 사용하여 표본평균을
  - 표준화한 것은 t 분포를 따름.
  - 단 하나의 분포가 아니라 자유도가 변함에 따라 분포가 달라짐
  - 자유도가 30 이상이면 표준정규분포 N(0 ,1)에 근사



여러 가지 표본 분포들

❑ F 분포

- 두 정규 모집단의 분산 비교에 대한 추론에 사용하는 분포
- 두 모 분산의 비에 대한 통계적 추론, 분산분석 등에서 유용하게 활용
- 비대칭 분포이며 여러 가지 자유도에 대한 분포 군이 존재
- 자유도가 커질수록 정규분포의 형태와 유사



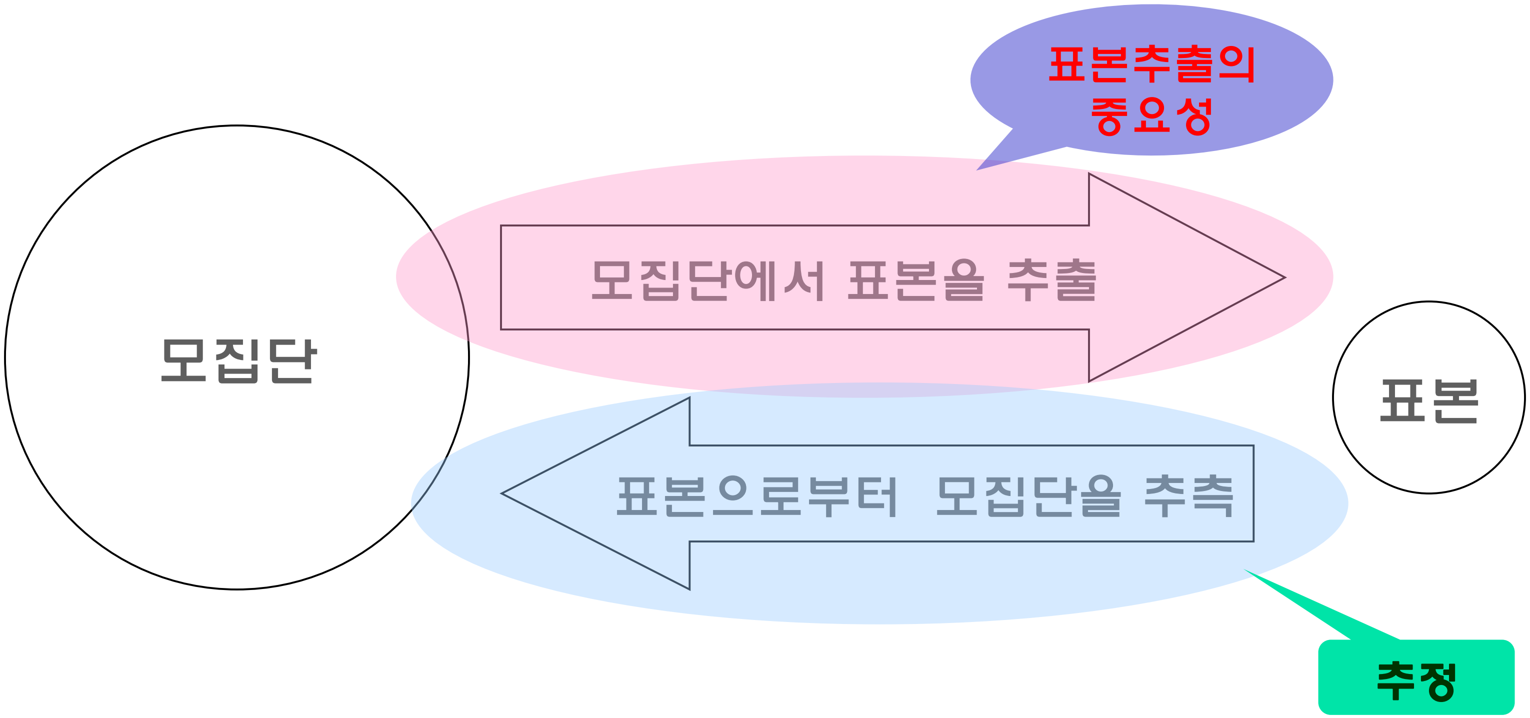
[연구 데이터 분석]

제2장　가설검정과 추정

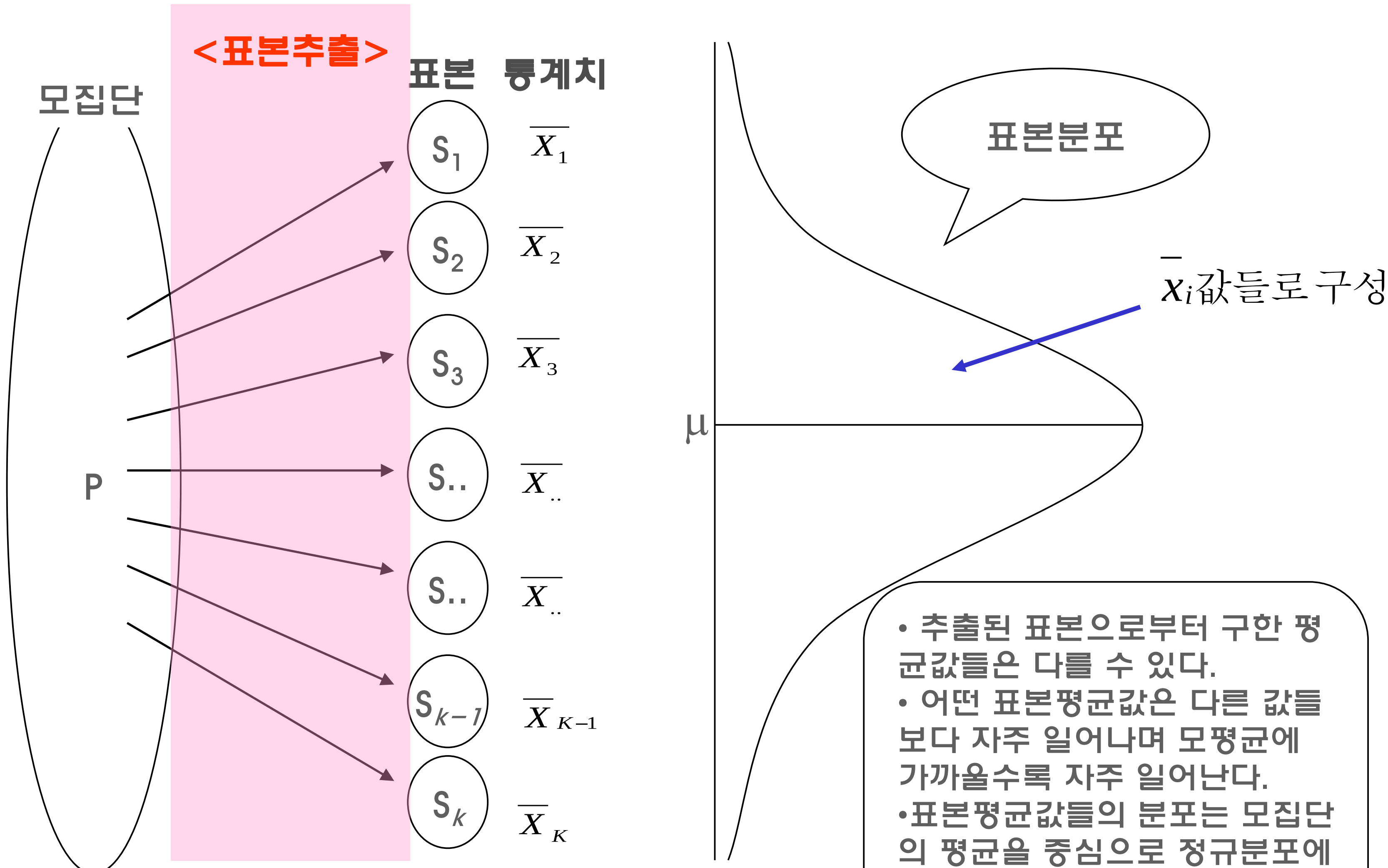
2.1 추론통계 개요
2.2가설검정
2.3 점추정과 구간추정

추정이란?

- 모든 자료를 조사할 수 없는 경우 표본에서 얻은 결과를 이용하여 모집단을 추측
- 모수(모집단 특성치) 추정에 확률이 핵심적인 역할
- **표본은 모집단을 대표할 수 있어야 함**



표본분포



- 추출된 표본으로부터 구한 평균값들은 다를 수 있다.
- 어떤 표본평균값은 다른 값들보다 자주 일어나며 모평균에 가까울수록 자주 일어난다.
- 표본평균값들의 분포는 모집단의 평균을 중심으로 정규분포에 접근한다.

점추정과 구간추정

점추정

- 표본으로부터 구한 통계치를 이용하여 모수를 특정한 값으로 추정(측)
- 구체적인 값으로 추측하지만 확률(가능성)에 대한 개념은 전무
- 모수에 대한 구체적인 가설이 있는 경우 : 점추정법을 사용
- 만약 $\mu = 10; \bar{x} = 9.9$ 라면?

구간추정

- 모수를 추측하는데 통계량의 분포를 이용, 통계치에 오차한계를 더하거나 빼서 모수가 들어있을 것으로 예상되는 구간을 제시
- 구체적인 가설을 가지고 있지 아니하고 표본 정보로부터 모수를 추측하고자 할 때 사용

신뢰구간의 추정

중학 수학의 경우 : 참값의 범위

근사값±오차한계 ⇒ 근사값−오차한계≤참값<근사값+오차한계

통계학의 경우 : 모평균(참값)에 대한 95% 신뢰구간 추정

- 모평균을 모르는 경우 표본평균을 이용하여 신뢰구간 추정

$$\overline{X} \pm t_{.05} \cdot s_{\overline{X}} \quad \Rightarrow \quad \overline{X} - t_{.05} \cdot s_{\overline{X}} < \mu < \overline{X} + t_{.05} \cdot s_{\overline{X}}$$

$$\overline{X} \pm t_{.01} \cdot s_{\overline{X}} \quad \Rightarrow \quad \overline{X} - t_{.01} \cdot s_{\overline{X}} < \mu < \overline{X} + t_{.01} \cdot s_{\overline{X}}$$

신뢰구간 추정

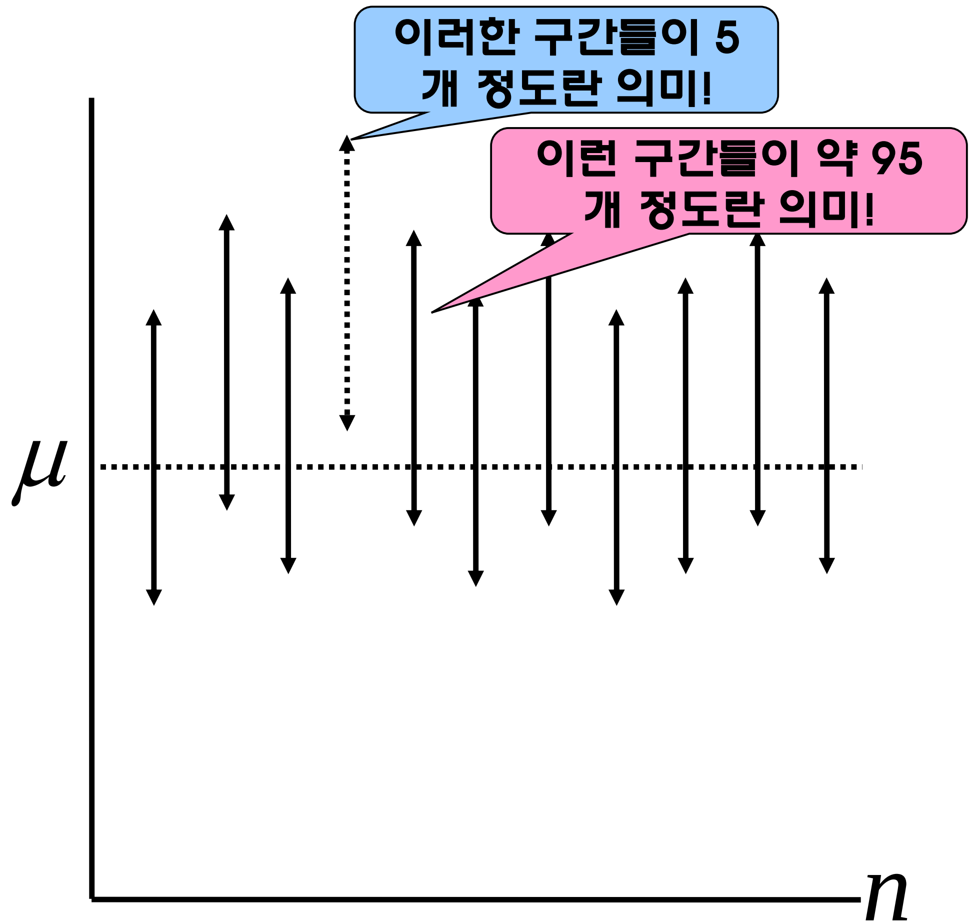
예)

- 모평균 μ 에 대한 95% 신뢰구간 : $(\bar{x} - d, \bar{x} + d)$

- 95% 신뢰수준의 의미 :

만일 크기가 30인 표본을 같은 방법으로 100번 추출하여(3,000 개체가 추출됨)

각 표본으로부터 100개의 신뢰구간을 구하면 그 중 95개 정도의 구간이 모수 μ를 포함함을 의미



신뢰구간 추정 예

예) 2002년 12월 대통령선거 예측조사결과(SBS)

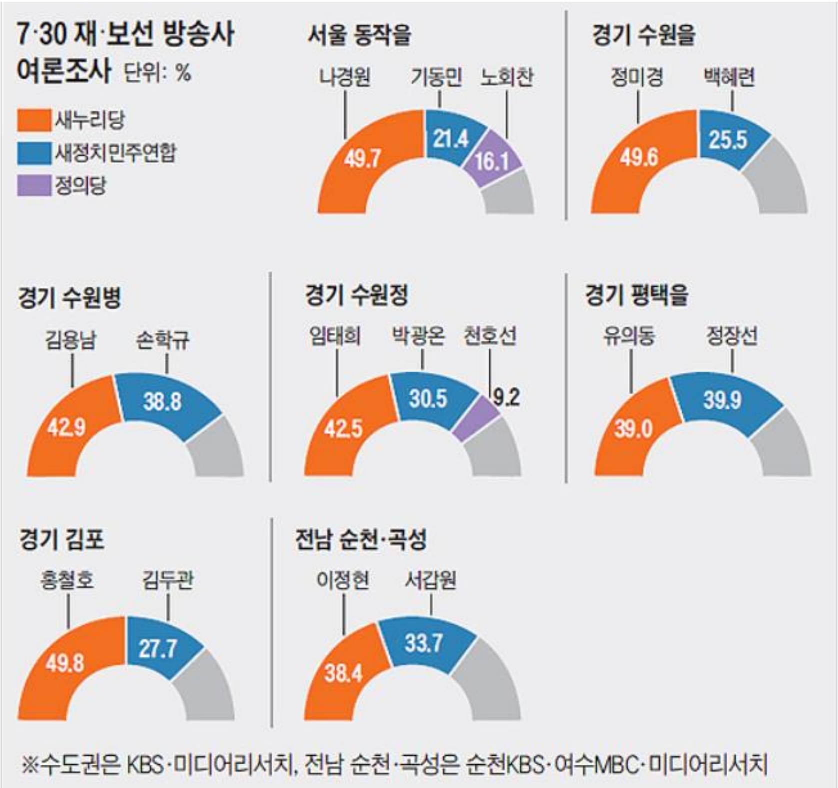
1. 점추정

- 나경원 후보 예측득표율 : 49.7%
- 기동민 후보 예측득표율 : 21.4%
- 오차의 한계(오차범위, 표본오차) : 3.1 %p

2. 구간추정

- 나경원 후보에 대한 95% 신뢰수준의 구간 : (49.7−3.1, 49.7+3.1)
- 기동민 후보에 대한 95% 신뢰수준의 구간 : (21.4−3.1, 21.4+3.1)
- 두 후보에 대한 예측범위가 겹치면

보합세로 판단



7·30 재·보선이 치러지는 수도권 6곳에 대해 KBS가 미디어리서치에 의뢰해 지난 22~23일 실시한 여론조사에서, 새누리당이 4곳에서 앞섰고 2곳은 새정치민주연합과 접전을 벌이는 것으로 나타났다.

서울 동작을은 새누리당 나경원 후보 49.7%, 새정치연합 기동민 후보 21.4%, 정의당 노회찬 16.1% 등의 순이었다. 기 후보와 노 후보가 단일화할 경우를 전제로 한 나 후보와의 양자 대결은 조사하지 않았다. 기 후보와 노 후보의 지지율을 합해도(37.5%) 나 후보(49.7%)와의 차이가 12.2%포인트였다.

여론조사 공표 시한(24일)을 앞두고 실시된 이번 여론조사는 지역별로 유권자 700명을, 전남 순천·곡성 조사는 유권자 1005명을 대상으로 실시했다. 유선전화 RDD(임의번호걸기)로 실시한 각 조사의 오차 범위는 95% 신뢰 수준에서 지역별로 ±3.1~3.7%포인트다.

고민 방법 ---- 보수적 입장에서 고민하기로 함.

. 기존 입장과 주장하고자 하는 입장이 부딪힌다면 아주 특별한 이유가 없는 한 기존입장을 생각하는 경향. (새 주장을 받아들이는 데는 매우 인색함)

. 항상 그렇지는 않음 – 보일 수 없거나, 힘든 것을 기존 입장으로 한다

무죄 와 유죄: 같다 와 다르다: 독립이다 와 독립이 아니다:

정규분포를 따른다 와 따르지 않는다:

. 가설검정이론 때문에 "= " 는 반드시 귀무가설에만 포함된다.

H_0 : $\mu = 450$ H_A : $\mu > 450$

H_0 : 독립이다. H_A : 독립이 아니다

H_0 : $p_1 \le p_2$ H_A : $p_1 > p_2$

2.1 가설검정

가설검정(Hypothesis Testing)

평균에 대한 가설

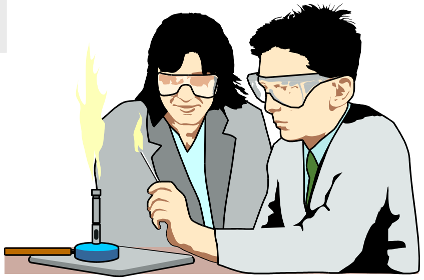
$H_0: \mu \geq 13.6$

$H_1: \mu < 13.6$

표준편차에 대한 가설

$H_0: \sigma_A \leq \sigma_B$

$H_1: \sigma_A > \sigma_B$



가설검정의 기본

귀무가설(H_0): 종래에 믿어오던 사실이나 보편적인 주장

대립가설(H_1): 새로운 주장

- 귀무가설이 '참' 이라고 가정하고, 그런 다음 이 가설을 채택하거나 기각할 수 있는 신빙성 있는 증거를 데이터에서 찾는다

- 귀무가설을 기각한다면, 대립가설을 채택한다

예제 : 통계적 의사결정

H_o 를 기각할 것인지 아닌지를 결정할 때, 2가지 의사결정 실수를 할 수 있다

		진 실	
		H_o 참	H_o 거짓
당신의 결정	H_o 채택	맞음	제2종 과오 (β 위험)
	H_o 기각	제1종 과오 (α 위험)	맞음

예제 : 재판

배심원의 결정 \ 진 실	실제로 무죄	실제로 유죄
무죄이다	맞음	제2종 과오 (β 위험)
유죄이다	제1종 과오 (α 위험)	맞음

결과 : 범죄자가 자유의 몸이 된다

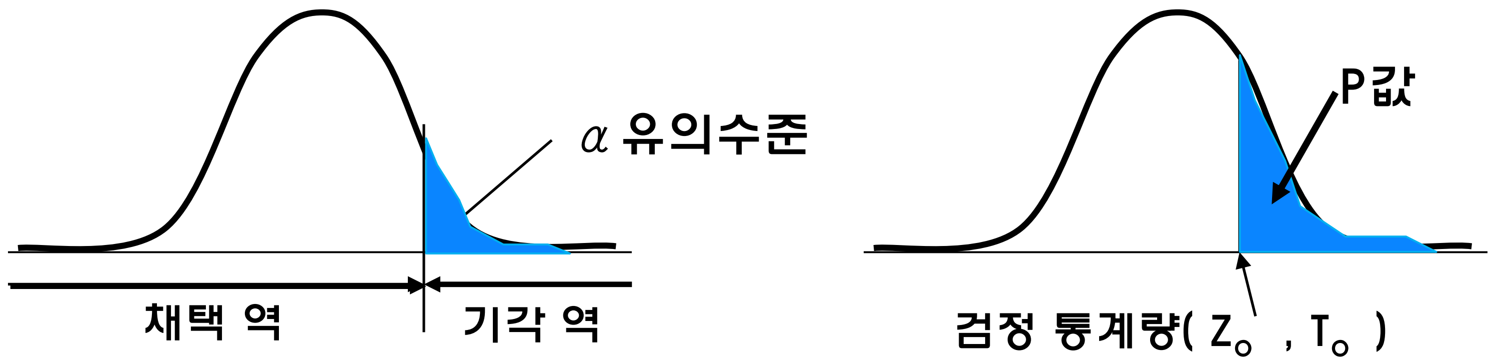
결과 :죄 없는 사람이 감옥에 간다

2.1 가설검정

- **유의수준(α)**
  – 귀무가설(H_0) 참인데도 불구하고 H_0을 기각할 확률의 최대 허용한계

- **기각 역**
  – 귀무 가설 (H_0)을 기각하는 영역
  – 검정 통계량이 기각 역에 있으면 귀무 가설 (H_0)을 기각하고 대립가설을 채택함

- **P값 (유의 확률)**
  – 정의 :



2.1 가설검정

검정 추정 개념

검 정

귀무 가설 (H_0)
대립 가설 (H_1)

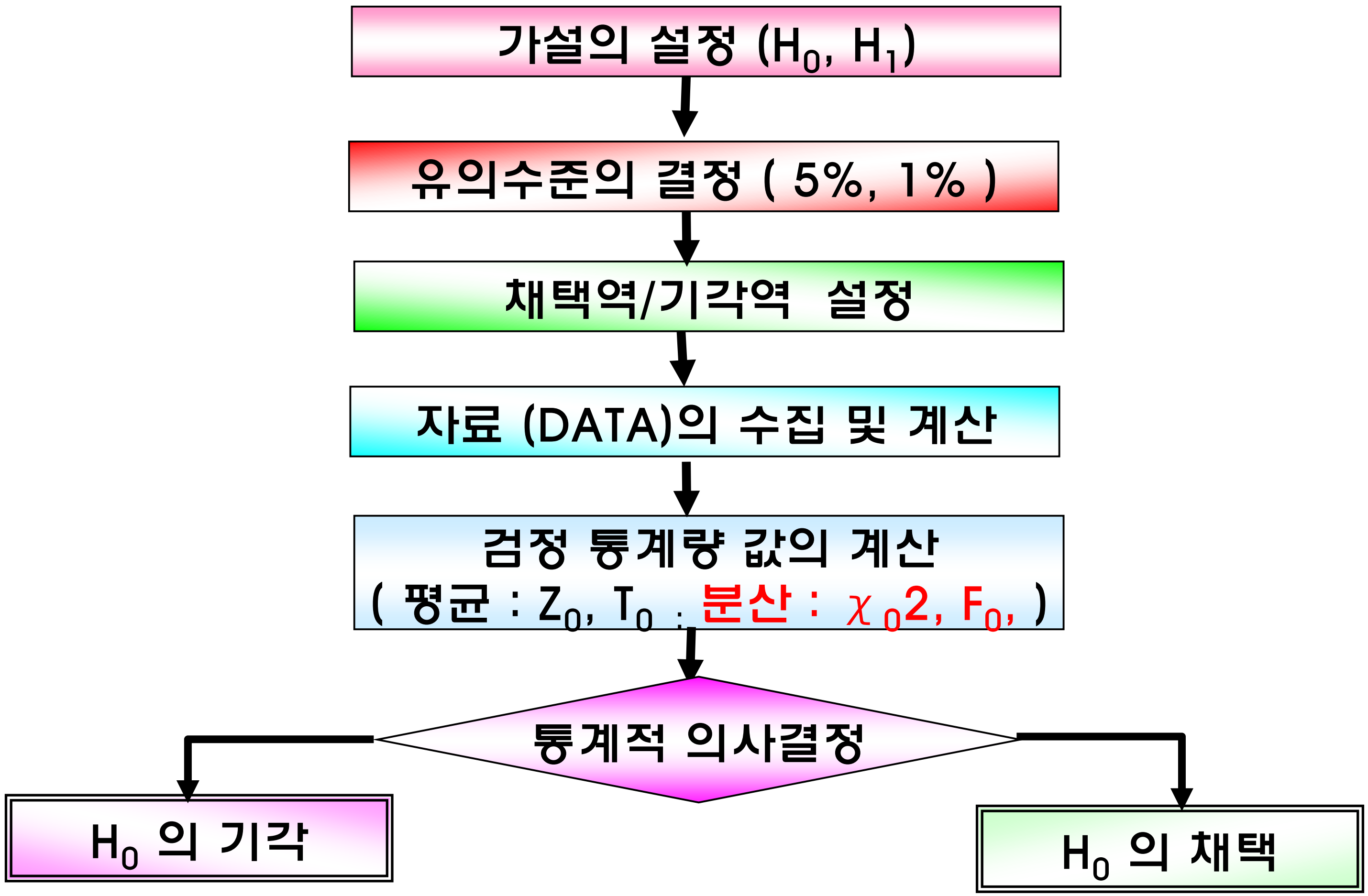
H_0 채택 또는 기각
판단으로 모집단의 평균과 분산은 같다, 또는 다르다 및 커졌다 /작아졌다 등으로 의사결정

추 정

점 추정 :
모평균 값은 얼마 일 것이다

구간 추정 :
모평균 값은 최소 얼마 내지 최대 얼마 일 것이다

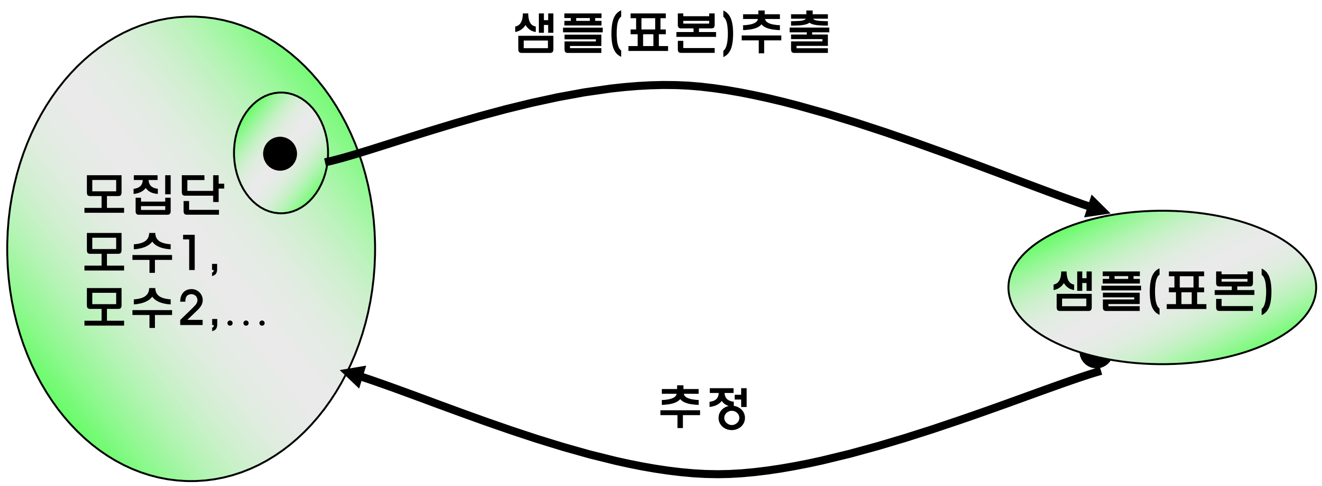
2.1 가설검정



※ 여기서 " o " 는 observed 즉 " 계산된 값 " (관측한 값)을 말하며 이를 검정 통계량 값이라 한다

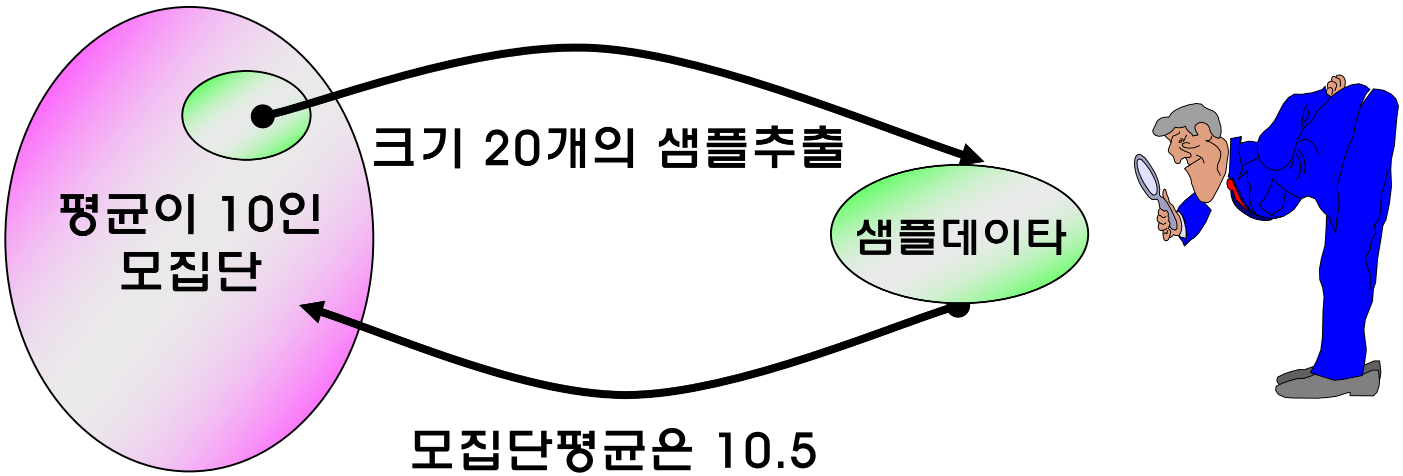
2.2 점추정과 구간추정

- 모수(Parameter) : 모집단의 분포 모양을 결정하는 수치적측도
  (모평균, 모분산, 모표준편차, 모공분산, 모상관계수 등과 같이 모집단의 특징을 나타내는 대표 값)



- 추정의 종류
  - **점 추정** : 모수의 추정치가 하나의 값(점)으로 주어지는 추정
  - **구간추정** : 모수의 추정치가 구간으로 주어지는 추정

- 관심 있는 모집단의 모수를 하나의 값으로 추정하는 방법
- 일반적으로 모집단의 모수 중에서 중요한 것들로는 평균, 분산, 표준편차 등이 있음



구 분	모집단(N)	Sample(n)
평균	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_1$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
분산	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- **관심 있는 모집단의 모수를 구간으로 추정하는 방법**

- 구간 추정의 예

평균(μ) 구간추정	$\bar{x} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$ Known(기지) σ	$\bar{x} \pm t_{\alpha/2}(n-1) \dfrac{s}{\sqrt{n}}$ Unknown (미지) σ
분산(σ2) 구간추정	$\dfrac{(n-1)s^2}{x^2_{\alpha/2}(n-1)} \le \sigma^2 \le \dfrac{(n-1)s^2}{x^2_{1-\alpha/2}(n-1)}$	

2.2 점추정과 구간추정

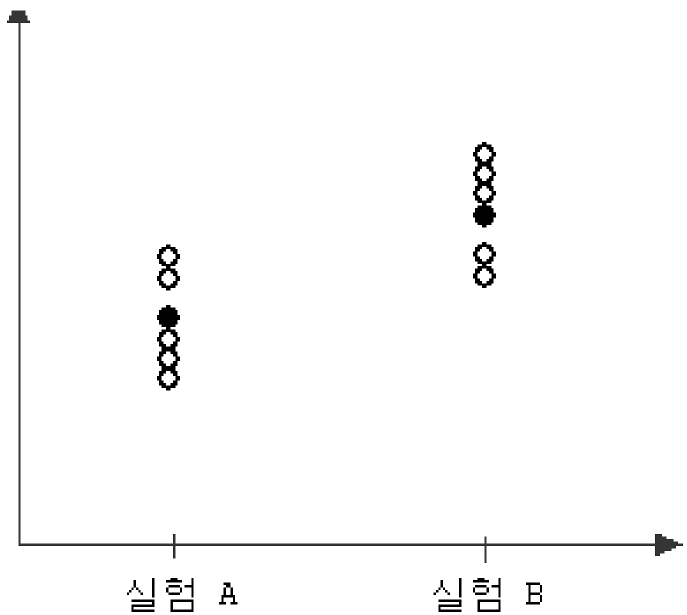
평균		검정통계량	구간추정
σ 기지 (Z)	모집단 1개	$Z_0 = \dfrac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}}$	$\bar{X} \pm Z_{\alpha/2} \dfrac{\sigma_0}{\sqrt{n}}$
	모집단 2개	$Z_0 = \dfrac{\bar{X}_A - \bar{X}_B}{\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}}$	$(\bar{X}_A - \bar{X}_B) \pm Z_{\alpha/2} \sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}$
σ 미지 (T)	모집단 1개	$T_0 = \dfrac{\bar{X} - \mu_0}{s / \sqrt{n}}$	$\bar{X} \pm t_{\alpha/2}(\varphi) \dfrac{s}{\sqrt{n}}$
	모집단 2개 $(\sigma_A^2 = \sigma_B^2)$	$T_0 = \dfrac{\bar{X}_A - \bar{X}_B}{\sqrt{\left(\dfrac{1}{n_A} + \dfrac{1}{n_B}\right) s_p^2}}$	$(\bar{X}_A - \bar{X}_B) \pm t_{\alpha/2}(\phi) \sqrt{\left(\dfrac{1}{n_A} + \dfrac{1}{n_B}\right) S_p^2}$
	모집단 2개 $(\sigma_A^2 \neq \sigma_B^2)$	$T_0 = \dfrac{\bar{X}_A - \bar{X}_B}{\sqrt{\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}}}$	$(\bar{X}_A - \bar{X}_B) \pm t_{\alpha/2}(\phi^*) \sqrt{\left(\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}\right)}$

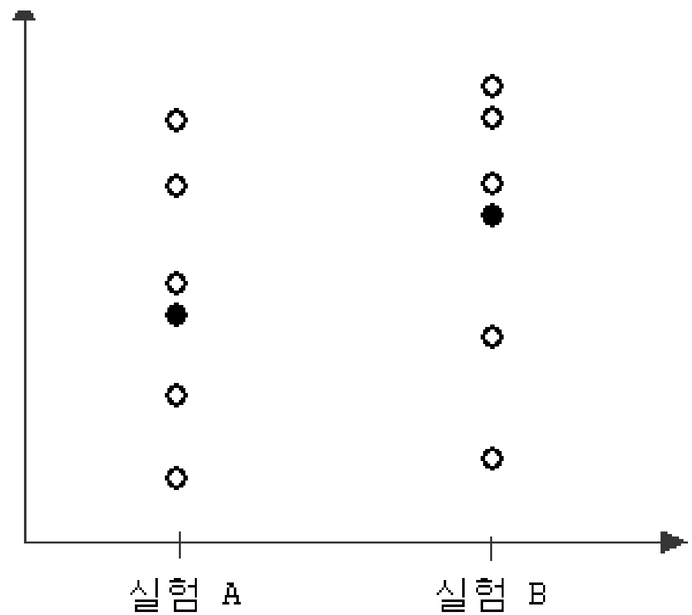
[연구 데이터 분석]

제3장　비교분석

3.1 비교분석 개요
3.2 단일모집단 평균 t-test
3.3 두 모집단 평균 비교 t-test
3.4 분산분석

예) 흡연집단과 비흡연집단의 폐암 발생률의 비교(차이)
두 치료약(치료방법)에 따른 치료율 비교(차이)
두 회사의 가전제품에 대한 선호도 비교(차이)





3.2 단일 모집단 평균 t-test

$$H_0 : \mu_1 = \mu_2 \Leftrightarrow H_0 : \mu_1 - \mu_2 = 0$$

세 가지 방향

1) 두 집단의 데이터가 서로 연관 : paired T−test

2) 두 집단의 데이터가 서로 독립 : T−test

− 두 집단의 분산이 서로 같은지 여부에 따라 분석 방법이 달라짐

3) 두 집단의 분산을 알고 있느냐? 모르느냐?

모른다면 표본의 크기가 크냐 작으냐?

즉, 중심극한정리를 사용할 수 있느냐? 없느냐?

두 모집단의 혹은 성질이 서로 다른 두 집단의 평균비교 즉, 두 집단의 비교분석 시 주로 사용

$$X_1, X_2, \cdots, X_{n_1} \sim iid\ N(\mu_1, \sigma_1^2)$$
$$Y_1, Y_2, \cdots, Y_{n_2} \sim iid\ N(\mu_2, \sigma_2^2)$$

$\Rightarrow$ $\overline{X} - \overline{Y}$ **의 분포는?**

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

$$=> \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

그런데, σ_1^2, σ_2^2 **모른다면?**

$$\hat{\sigma}_1^2 = \sum_{i=1}^{n_1} (X_i - \overline{X})^2 / (n_1 - 1)$$

$$\hat{\sigma}_2^2 = \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2 / (n_2 - 1)$$

3.2 단일 모집단 평균 t-test

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

$$=> \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$\hat{\sigma}_1^2 = \sum_{i=1}^{n_1} (X_i - \bar{X})^2 / (n_1 - 1)$$

$$\hat{\sigma}_2^2 = \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 / (n_2 - 1)$$

2) 두 모분산이 같을 경우 $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sum (X_i - \bar{X})^2}{n_1 - 1} + \frac{\sum (Y_i - \bar{Y})^2}{n_2 - 1}}} \sim t(n_1 + n_2 - 2)$$

3.2 단일 모집단 평균 t-test

- 예)
  - 설비 A의 수명: 0.9, 2.2, 1.6, 2.8, 4.2, 3.7, 2.6
  - 설비 B의 수명: 1.4, 2.7, 1.8, 3.0, 3.2

  - A 수명 : 평균=2.57, 표준편차=1.144
  - B 수명 : 평균=2.42, 표준편차=0.7823

  - t=0.2548, p−값=0.8040
  - $t_{0.05}(10)=2.23$

  - 실습

3.2 단일 모집단 평균 t-test(SPSS)

(예제1) 다음 자료는 모 기업의 일간 전력최대 사용량을 정리한 자료이다. 공휴일 여부에 따른 최대수요값의 차이가 있는지 분석하여라.

요일	공휴일여부	최대수요	최소기온	최대기온	평균풍속	최대풍속	강수량
토	1	24226	1.8	11.1	2.4	5.1	0
일	1	26027	1.5	9.5	3	7.2	0
월	2	32513	-2.2	4.1	2.6	6.3	0
화	2	34079	-2.9	7.5	1.5	3.7	0
수	2	34118	1.4	9	2.7	6	10.8
목	2	34413	-0.5	8.7	3.3	7.7	5.1
금	2	34604	-7.4	-0.4	3.1	8.4	0
토	1	32552	-5	2.9	2	5.3	0
일	1	28659	-3	2.2	1.7	4.8	0
월	2	34590	-2.3	5.2	2.6	6.6	0
화	2	34115	-4.5	7.1	2.5	6.1	0

(예제2) 두 종류의 사료가 젖소의 우유생산량에 미치는 영향의 차이를 조사하기 위해서 랜덤하게 8마리씩 A, B 두 그룹으로 나눈 후 A 그룹에는 사료 1을 B 그룹에는 사료 2를 주면서 3주일 동안의 우유생산량을 조사하였다. 두 종류의 사료가 우유 생산량에 미치는 영향이 다르다고 할 수 있는지를 유의수준 5%에서 검정하여라.

사료와 우유생산량

그룹A (사료1)	54	60	66	53	62	61	42	50
그룹B (사료2)	65	70	62	67	59	45	60	52

쌍체(대응, paired) 표본 검정

[예제]. S사에서는 직업훈련이 근로자들의 능률 향상에 효과가 있는지를 알아 보고자 한다.

- **독립표본**

근로자	1	2	3	4	5	6	7	8	9	10	11	12	13	14
실시전	76		76			87	67				65		86	83
실시후		84		88	77			77	75	78		83		

- **쌍체 (대응, 짝 지어진) 표본**

근로자	1	2	3	4	5	6	7	8	9	10
실시전	76	60	85	58	91	75	82	64	79	88
실시후	81	52	87	70	86	77	90	63	85	83

- **두 실험 설계의 차이점은?**
- **짝 지어진 표본은 언제 사용하는가?**
  - **배제할 기타 변동요인이 존재할 때 즉, 근로자들간의 능률 산포가 클 때**

- **절차상 다른 점**
  - **근로자들간의 능률 산포를 배제하기 위해서, 각 근로자의 원래 데이터가 사용되지 않고 차이가 사용, 차이는 순수하게 직업훈련의 효과만을 반영**

쌍체(paired) 표본 검정

- 데이터의 차이 계산

근로자	1	2	3	4	5	6	7	8	9	10
실시전	76	60	85	58	91	75	82	64	79	88
실시후	81	52	87	70	86	77	90	63	85	83
차이	−5	8	−2	−12	5	−2	−8	1	−6	5

- 가설 설정

$H_0 : \mu_A = \mu_B$ v.s. $H_1 : \mu_A \neq \mu_B$

$$T_0 = \frac{\overline{D}}{\sqrt{s_D^2 / n}} \sim t(n-1)$$

- 검정 통계량의 값(T_0 = −0.79) 과 p−value 계산

$$T_0 = \frac{-1.6}{6.38\sqrt{10}} \qquad p-value : 0.448$$

- 의사결정 : 직업훈련 전후에 능률에 차이가 없다 라는 H_0 채택

(예제 3) 자동차의 휘발유에 사용하는 첨가제가 주행거리에 영향을 미치는지 알아보고자 한다. 다섯 종류의 새 차에 대하여 같은 종류의 차 두 대 중에서 한 대를 랜덤하게 택하여, 첨가제를 사용하고 다른 한 대에는 첨가제를 사용하지 않고서 같은 운전자가 같은 장소에서 운전한 결과 다음과 같은 자료를 얻었다. 첨가제를 사용하는 경우 주행거리에 차이가 있다고 할 수 있는지 유의수준 5%에서 검정하여 보자.

휘발유 1ℓ 당 주행거리 (단위 km)

차의 종류	1	2	3	4	5
첨가제를 사용한 경우	11.8	13.9	16.3	11.6	8.4
첨가제를 사용 안한 경우	11.4	13.1	16.1	10.9	8.3

3.4 분산분석

- 일원분류 분산분석(one-way ANOVA)

  : 독립(설명)변수의 개수가 한 개

- 다원분류 분산분석(multi-way ANOVA)

  : 독립(설명)변수의 개수가 두 개 이상

- 일변량 분산분석(univariate ANOVA)

  : 반응변수의 개수가 한 개

- 다변량 분산분석(multivariate ANOVA)

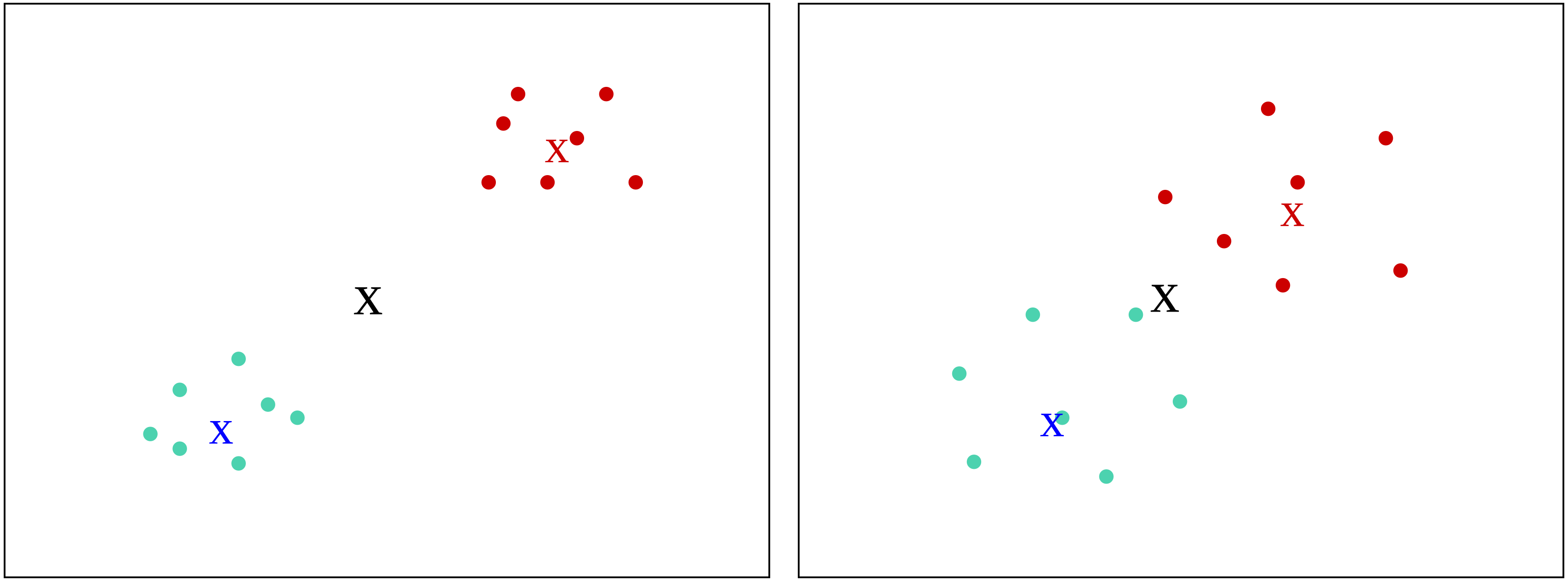
  : 반응변수의 개수가 두 개 이상

- 공분산분석(Analysis of Covariance)

  : 설명변수에 연속형인 공변량(covariate)이 포함되어 있는 경우

x : 전체평균
x : Group1 평균
x : Group2 평균

1) 원리



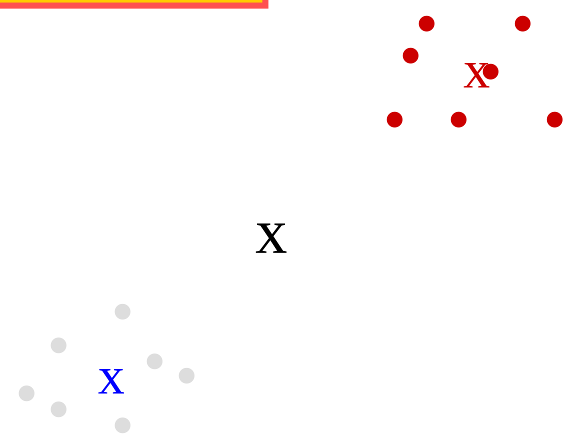
1) 원리

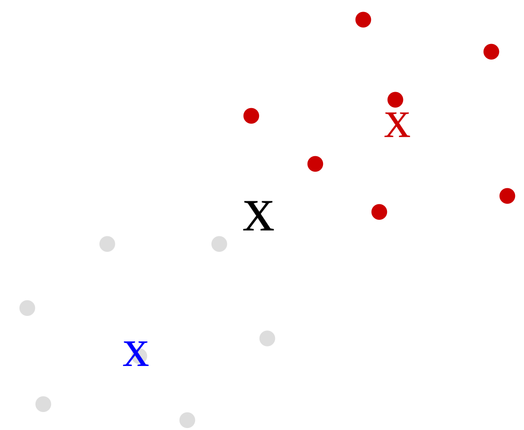
x : 전체평균
x : Group1 평균
x : Group2 평균

$$\sum(y_{g_2}-\bar{y}_{g_2})^2=31$$

그룹1내 변동

$$\sum(y_{g_2}-\bar{y}_{g_2})^2=78$$





$$\sum(y_{g_1}-\bar{y}_{g_1})^2=25$$

그룹2내 변동

$$\sum(y_{g_1}-\bar{y}_{g_1})^2=70$$

$$\sum_{i=1}^{2}(\bar{y}_{g_i}-\bar{y})^2=61$$

그룹간 변동

$$\sum_{i=1}^{2}(\bar{y}_{g_i}-\bar{y})^2=29$$

2) 기본모형

- **자료구조**

	그룹1	그룹2	…	그룹k
	y_{11}	y_{21}	…	y_{k1}
	y_{12}	y_{22}		y_{k2}
	$\vdots$	$\vdots$		$\vdots$
	y_{1n_1}	y_{2n_2}	$\vdots$	y_{kn_k}
평균	$\bar{y}_{1\bullet}$	$\bar{y}_{2\bullet}$		$\bar{y}_{k\bullet}$

총평균 $\bar{y}$

- 모형

$$Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i=1, \cdots, k \quad j=1, \cdots, n$$

여기서, m: 총평균, α_i : i번째 처리효과, ε_{ij} : 오차항

3) 총 변동의 이해

❖ **총 편차의 분해**

$$\mathbf{y}_{ij} - \bar{\mathbf{y}} = (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i}) + (\bar{\mathbf{y}}_{i} - \bar{\mathbf{y}})$$

- **총 변동의 분해**

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(\mathbf{y}_{i} - \bar{\mathbf{y}})^{2} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{\mathbf{y}}_{i} - \bar{\mathbf{y}})^{2} + \sum_{i=1}^{k}\sum_{j=1}^{n}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i})^{2}$$

전체제곱합(TSS) = 처리제곱합(SST) + 잔차제곱합(SSE)

4) 분산분석 표(ANOVA Table)

분산의 요인	제곱합	자유도	평균제곱	분산비
처리(Treatment)	SST	k-1	MST	F=MST/MSE
오차(Error)	SSE	N-k	MSE	
전체(Total)	TSS	N-1		

❖ **F-검정**

: k개 집단간의 반응변수의 평균차이가 있는가를 검정

귀무가설 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

검정 통계량 : $F = \dfrac{MST}{MSE} = \dfrac{SST/(k-1)}{SSE/(N-k)}$

5) 다중비교

- 다중비교의 필요성

– T–검정은 제 1종오류(type I error)를 크게한다.

P(제 1종오류) = $1-(1-\alpha)^2 = (1-0.05)^2 = 1-0.86 = 0.14$

- 다중비교 방법들

– LSD, TUKEY, DUNCAN, BON, SCHEFFE, WALLER

TUKEY : 다중비교에 있어서의 실제 유의수준은 α 보다 약간 작게 된다. 어느 두 수준의 평균값의 차이가 근소할 때 이를 민감하게 검출하지 못한다는 단점이 있다.

DUNCAN : 두 평균값의 차이를 검출하는데 있어서 TUKEY의 방법보다 약간 더 민감하다.

연습문제

1. (예제1) 보험자료에 대하여 나이를 다음과 같이 3 그룹으로 나누어 각 그룹별로 보험가입금액과 월수입의 평균과 표준편차를 구하라.

그룹 1: 나이 35세 미만
그룹 2: 35 – 50세
그룹 3: 51세 이상

(참고: 분석–평균비교–집단별 평균분석 절차를 이용하기 바람)

2. 어떤 화학약품의 제조에 상표가 다른 2 종류의 원료가 사용되고 있다. 각 원료에서 주성분 A의 함량은 다음과 같다. 단, 함량들은 정규분포를 따른다고 가정한다. 이 두 원료의 주성분 A의 함량이 다른가를 분석하라.(화학제품 함량)

상표 1	80.4	78.2	80.1	77.1	79.6	80.4	81.6	79.9	84.4	80.9	83.1
상표 2	80.1	81.2	79.5	78.0	76.1	77.0	80.1	79.9	78.8	80.8	

연습문제

3. 특정 피임약이 사용자의 혈압을 저하시키는가 조사하고자 한다. 이를 위해 15 명의 부인들을 대상으로 평상시 혈압을 측정한 뒤 이들에게 이 피임약을 일정 기간 사용하게 한 후 이들의 혈압을 다시 측정한 결과를 기록했다. 얻어진 데이터는 다음과 같다. 피임약 복용이 혈압에 영향을 주는가 분석하라.

부인	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
사용 전	70	80	72	76	76	76	72	78	82	64	74	92	74	68	84
사용 후	68	72	62	70	58	66	68	52	64	72	74	60	74	72	74

4. 어떤 화학조미료의 개발연구를 행한 결과 방법1과 2중에 하나를 선택하기로 하였다. 원료 10 로트에 대하여 pilot plant에서 실험결과 다음 수확량의 데이터(kg)를 얻었다.

방법1	72	70	84	78	81	75	82	64	79	88
방법2	81	52	80	70	86	77	80	63	75	83
차이	−9	18	4	8	−5	−2	2	1	4	5

(1) 수확량이 더 많은 방법은 무엇이냐? 어떤 검정을 실시하여야 하는가?

(2) 위의 검정을 Paired t−test로 하지 않고, 독립 t−test를 실시한다면 어떤 결과가 얻어지느냐? 검정하여 보아라.

(3) 방법1, 2에 의한 수확량 모평균의 95% 신뢰구간을 구하여 이들을 비교하여 보아라. 어떤 정보가 얻어지느냐?

[연구 데이터 분석]

1. 상관분석
2. 회귀분석
3. 단순회귀분석 예제
4. 중회귀분석
5. 중회귀분석 예제

제4장 상관 및 회귀분석

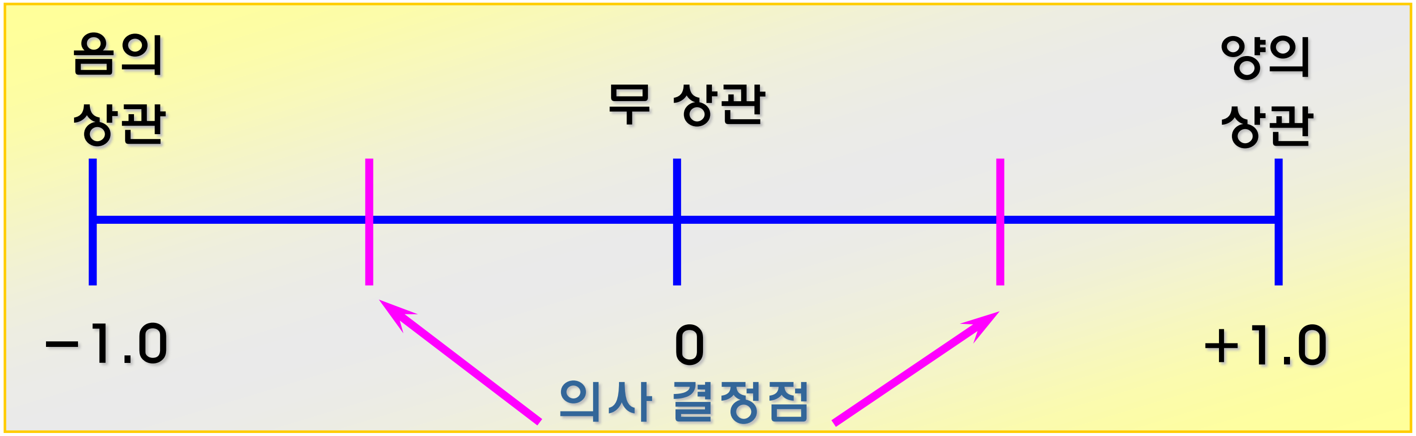
❑ 상관 회귀 분석

- 의 의
  - 변수(x_1)와 변수(x_2)사이 또는 X와 Y사이에
    - 얼마만큼의 관련성이 있는지 알아보고 – 상관분석
    - 함수관계를 도출하고 출력변수를 예측 – 회귀분석
- 분석 목적
  - 이들 간에는 얼마나 강한 관계가 있을까?
  - 이들 간에는 어떠한 관계식이 있을까?
- 관련성 확인(예)
  - 지능지수 vs 학업성적
  - 흡연량 vs 폐암발생률
  - 공정온도 vs 제품강도

1) 상관계수

– **필요성**

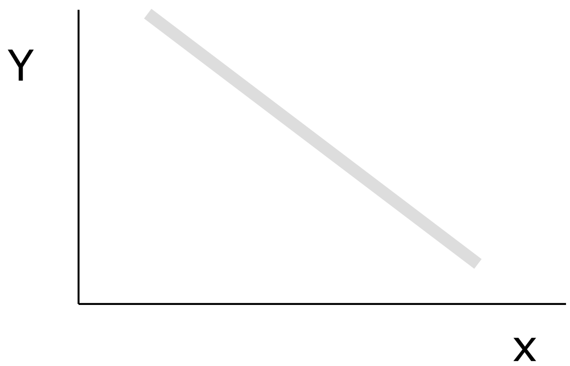
- 상관관계는 두 변수들 사이에 얼마만큼의 관련성이 있는지를 수치적으로 알아볼 수 있다.
- 두 변수 사이의 연관성의 강도는 상관계수(r)를 이용하여 계수화
- 보통 Pearson’s product moment 상관계수를 사용한다.



상관계수 (r) : 두 변수의 상호 의존관계를 양적으로 나타내는 척도

– 상관계수의 성질

- r 값이
  - (+) 이면 양의 상관관계
  - (–) 이면 음의 상관관계
  - 0 에 가까우면 상관관계 없음



r 이 −1에 근접

r 이 +1에 근접

- 상관 관계를 조사하기 위해서는 데이터 구조가 순서쌍으로 이루어진 이변량 데이터 구조가 요구된다.

– 모 상관계수 (Correlation Coefficient)

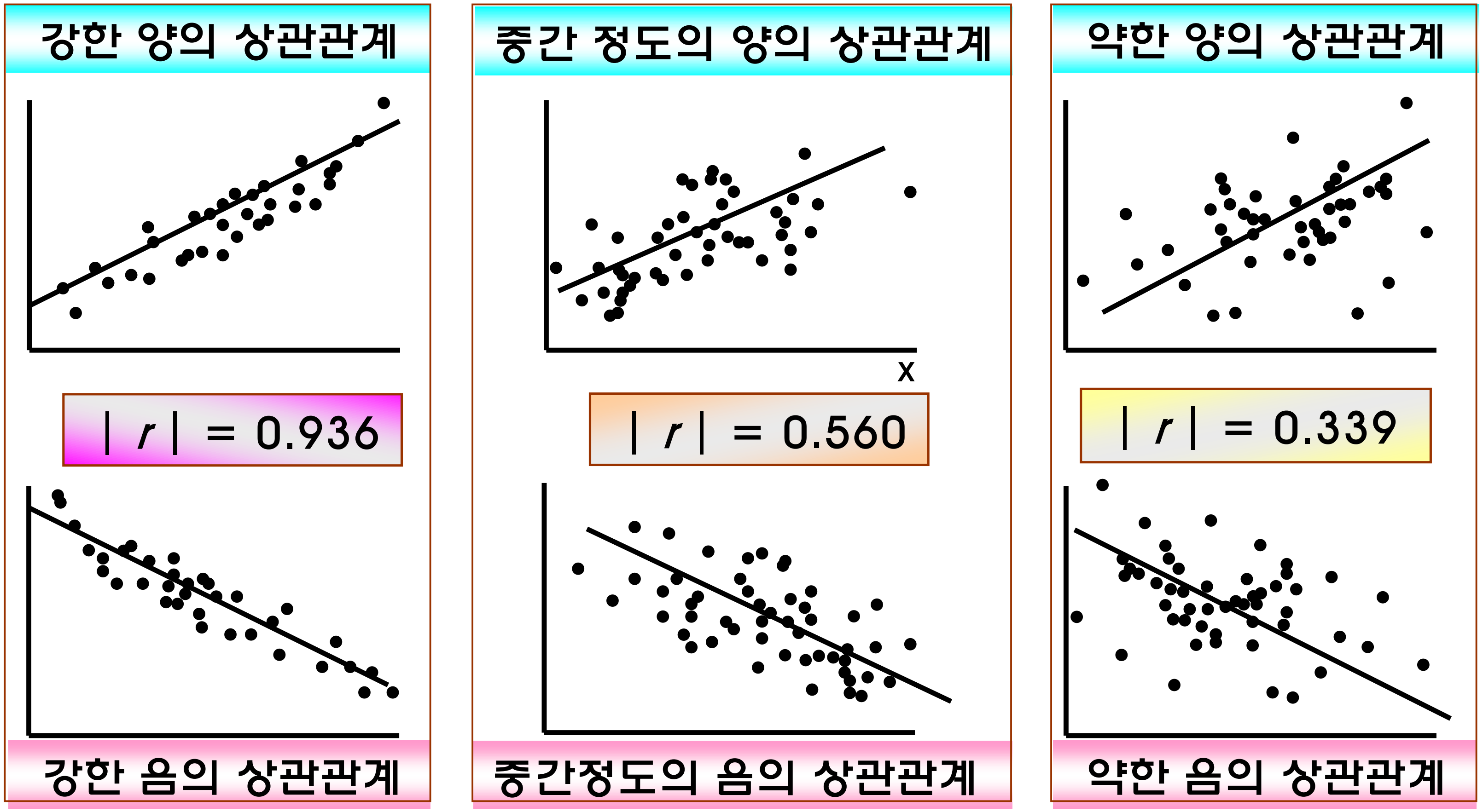
- 일반적으로 ρ 로 표시하며 그 범위는 $-1 \le \rho \le 1$이다.
- 그러나 ρ 의 정확한 값은 알 수 없다. 따라서 샘플로부터 추정한 값 표본상관 계수 r을 사용한다. r은 다음 식에 의해 구해지며, 언제나 $-1 \le r \le 1$ 이다.

표본상관 계수 공식

$$\hat{\rho} = r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

4.1. 상관분석

– 상관관계 유형



2) 상관분석의 함정

- Y와 X 간에 상관이 있다는 것을 입증했다 하더라도, 이것이 반드시 Y의 변동이 X의 변동에 의해서 초래되었다는 것을 의미하지는 않는다. X와 Y 모두에 변동을 초래하는 제3의 변수가 “숨어” 있을 수 있다.
- 두 변수 간에 관계가 있다는 결론이 인과관계를 의미하는 것은 아니다.
- 표본상관계수의 값이 “0”에 가깝다는 것은 두 변수 사이의 직선관계가 약하다는 뜻이지, 반드시 두 변수 사이에 관계가 없음을 뜻하는 것은 아니다.

>> 상관관계가 있다고 해서 반드시 인과관계가 있는 것은 아니다.

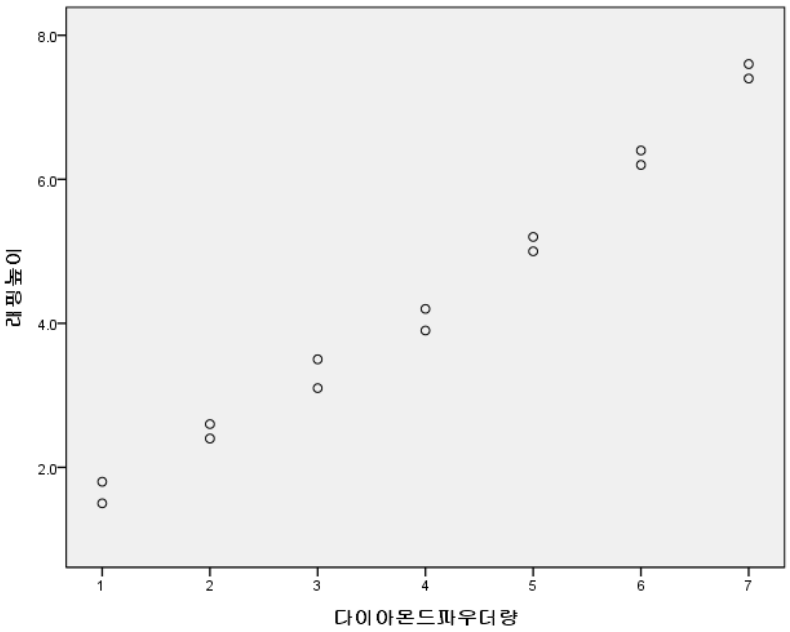
상관은 인과관계를 파악하는 것이 아니다 !

4.1. 상관분석

❑ M 제품의 면을 다듬기 위하여 Lapping을 하고자 한다.
Lapping시 Diamond powder를 사용하는데 Powder의 사용량에 따라 Lapping된 높이를 알고 싶어 한다. 이를 알아보기 위하여 여러 번의 실험을 하였는데, 이 자료의 산점도를 구해보고 표본 상관계수를 구하시오.
<래핑데이터.sav>

▷ 항상 데이터를 그래프 상에 타점하는 산점도 수행을 먼저 실시.

▷ 그런 다음, 선형 관계가 보이면 상관분석을 실시.



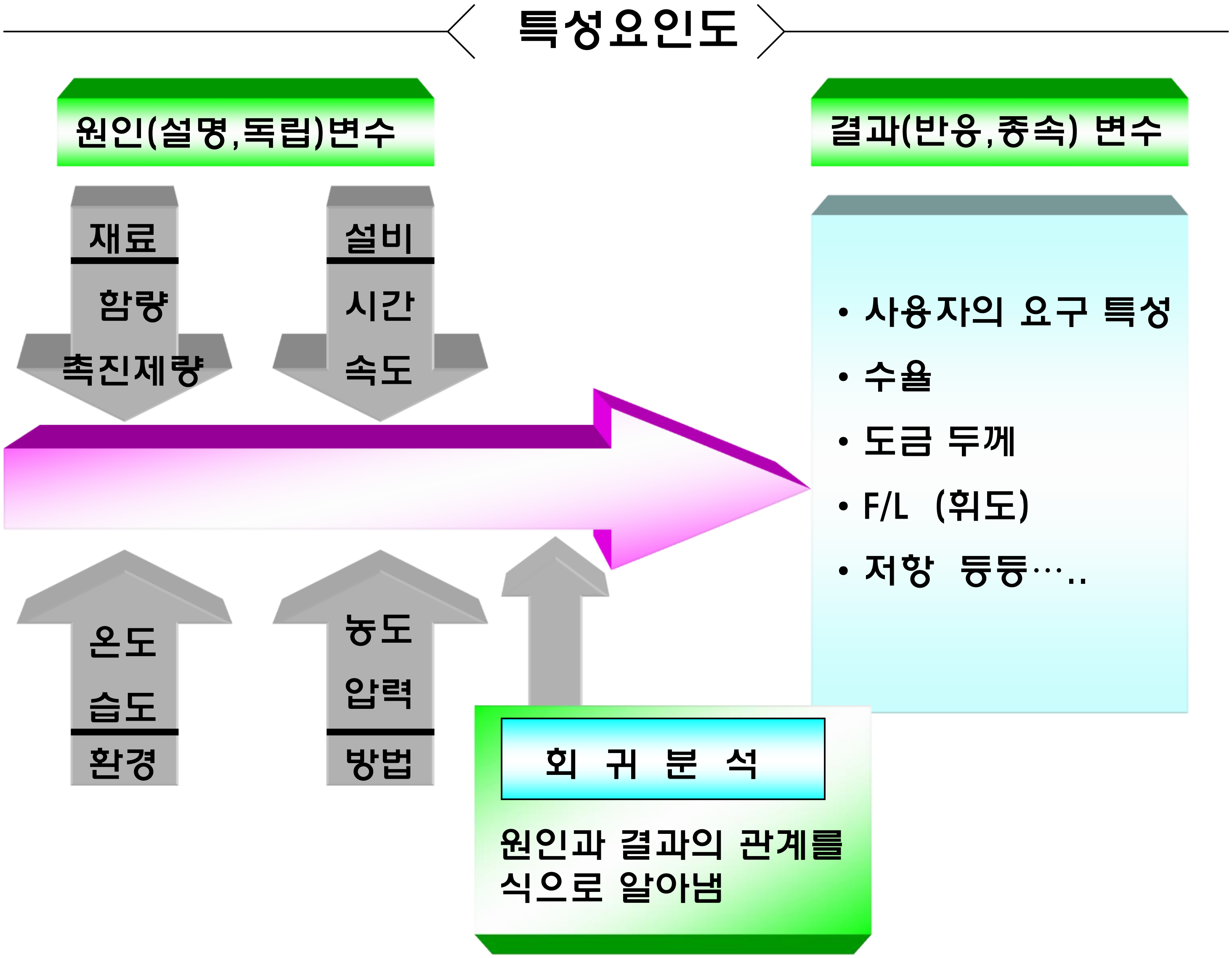
회귀분석이란?

– **필요성**
  - 회귀 분석은 입력변수(X)들이 출력 값(Y)에 미치는 영향을 예측하고자 할 경우에 그 관계를 **함수관계(회귀식)와 결정계수**로 나타내어 분석하는 방법론.
  - 이를 통해 출력 값(Y)에 어떤 인자가 얼마만큼의 영향을 미치는지 알아내어 우리가 원하는 출력 값을 얻기 위하여는 X를 어떤 수준으로 얼마만큼 관리 해야 되겠다는 정보를 알 수 있도록 해 줌

– **회귀 방정식**
  - 입력변수의 값을 사용해서 이에 상응하는 출력 값에 대한 예측을 할 수 있게끔 해 주는 예측 방정식이다.

– **결정계수 (기여율)**
  - R^2, 회귀 모형의 적합성 또는 총변동 중에서 회귀식에 의해 설명된 변동의 비율을 나타낸다.

4.2. 회귀분석



특성요인도
원인(설명,독립)변수
결과(반응,종속) 변수
재료
함량
촉진제량
설비
시간
속도
온도
습도
환경
농도
압력
방법
• 사용자의 요구 특성
• 수율
• 도금 두께
• F/L (휘도)
• 저항 등등…..
회 귀 분 석
원인과 결과의 관계를
식으로 알아냄

4.2. 회귀분석

종류	특징	모형
단순 회귀 (Simple Regression)	독립변수가 1개이며, 종속변수와의 관계가 직선이다	$Y = \alpha + \beta x + \varepsilon$
곡선회귀 (Curvilinear Regression)	독립변수가 1개이며, 종속변수와의 관계가 곡선이다	2차 곡선인 경우 : $Y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$ 3차 곡선인 경우 : $Y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$
중 회귀 (Multiple Regression)	독립변수가 k개($x_1, x_2, \cdots, x_k$)이며,종속변수와의 관계가 선형(1차 함수)이다.	$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$
다항회귀 (Polynomial Regression)	독립변수가 k개($x_1, x_2, \cdots, x_k$)이며,종속변수와의 관계가 1차 함수 이상이다. (단, k=1이면 2차 이상)	k = 2이고 2차 함수인 경우 : $Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$
비 선형회귀 (Nonlinear Regression)	회귀식의 모양이 미지의 모수 β_i의 선형관계로 이루어져 있지 않다.	예 : $Y = \alpha e^{-\beta x} + \varepsilon$

1) 단순회귀분석

– 하나의 독립변수(X)와 하나의 종속변수(Y)간의 관계를 직선 방정식화하여 나타내기 위한 방법.

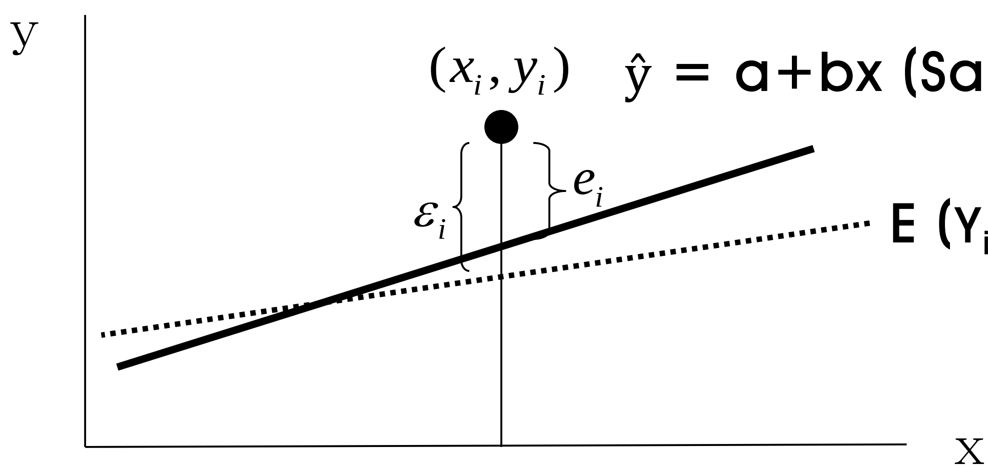
- Model

Independent & Identically Distributed

(독립이고 같은 분포를 따른다.)

$y_i = \alpha + \beta X_i + \varepsilon_i$ 여기서, $\varepsilon_i \sim N(0, \sigma^2)$

σ^2 : Unknown constant (미지상수)

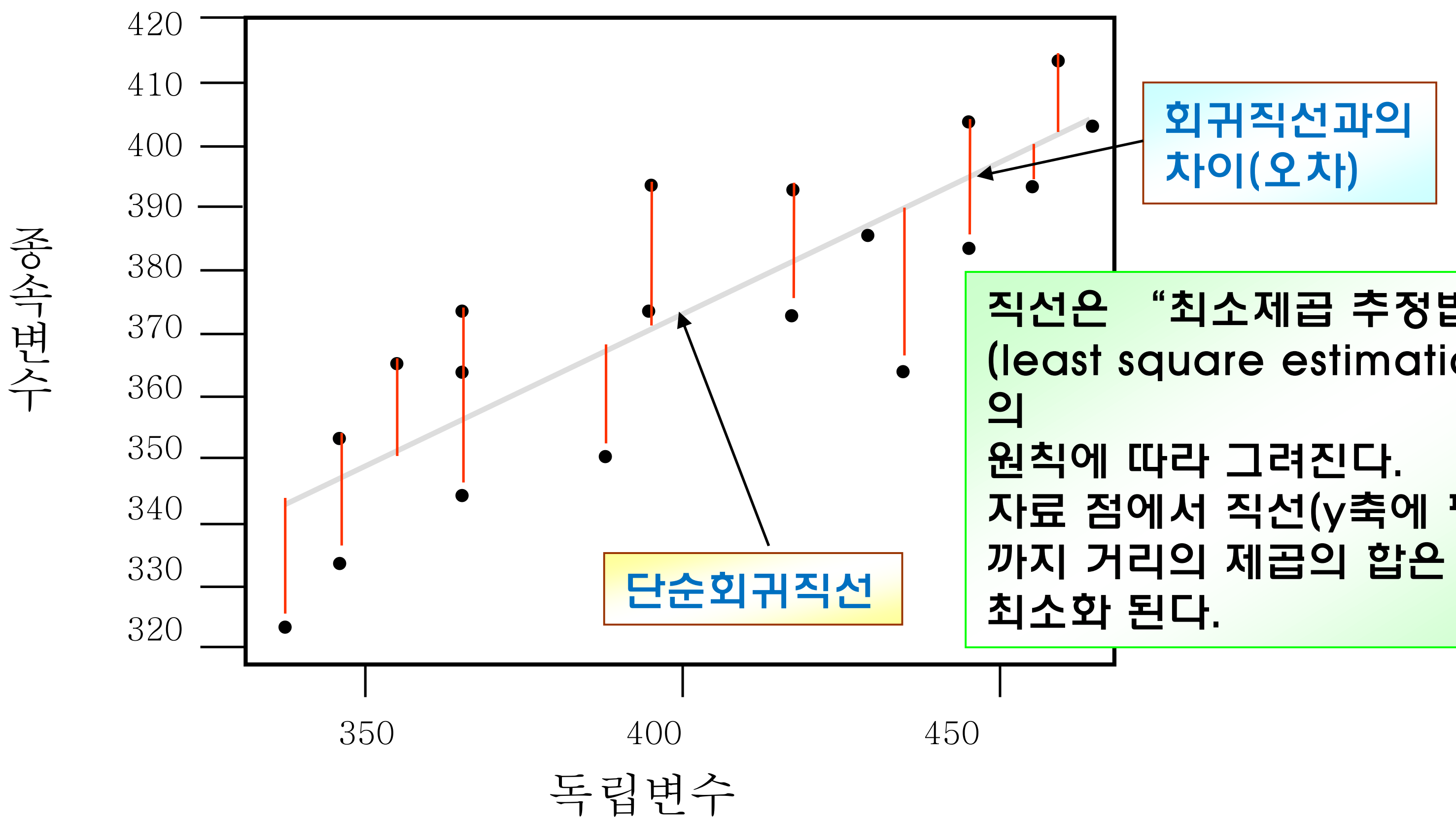


$\hat{y}$ = a+bx (Sample을 통해 추정한 직선)

E (Y_i)= $\alpha + \beta\,\chi$ (미지의 참 직선)

최소제곱법에 의한 단순회귀

– 오차 제곱 합을 최소로 하는 추정방법

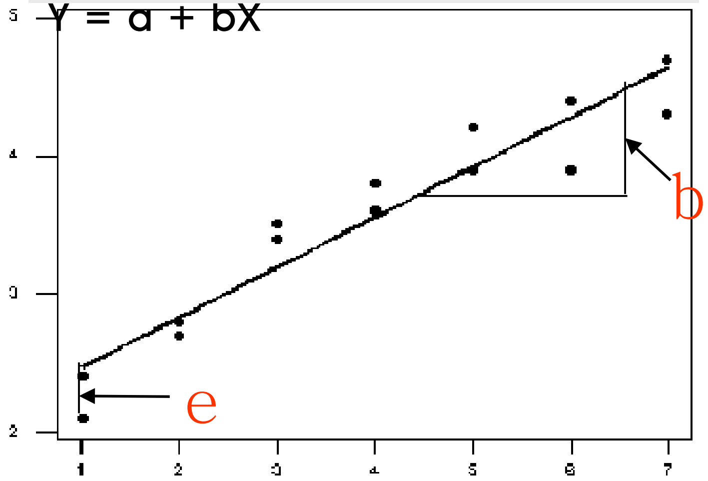


직선은 "최소제곱 추정법 (least square estimation)"의
원칙에 따라 그려진다.
자료 점에서 직선(y축에 평행)까지 거리의 제곱의 합은 최소화 된다.

❑ 회귀 방정식

Scatter Plot Y vs.X with Fitted Line

Y = a + bX

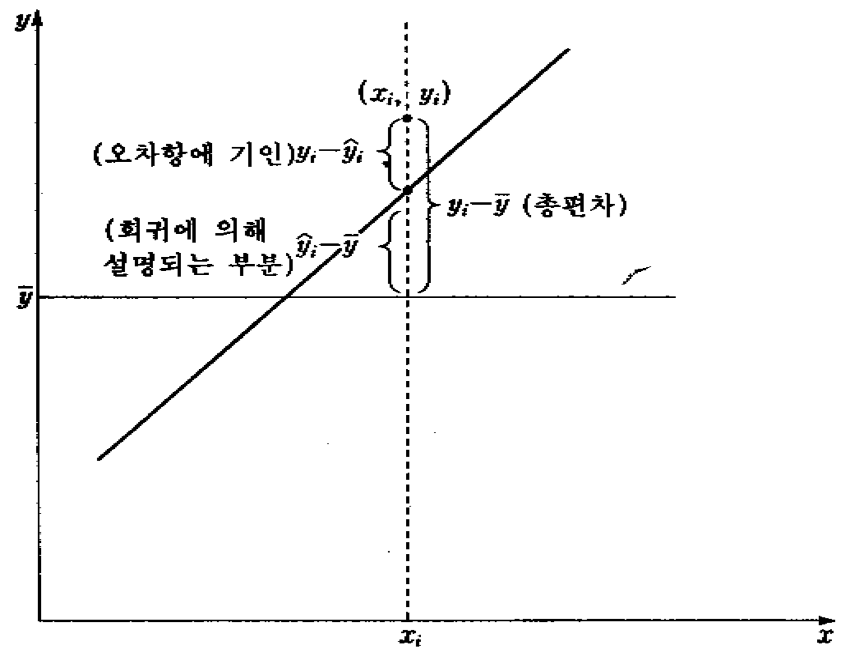


- 직선의 방정식은 Y = a +bX
- a는 Y–절편(x=0에서)이고 b는 기울기임
- 실제 자료 점들과 직선 사이의 차이는 잔차(residuals(e))라고 불린다.

2) 변동의 분해

- 단순 회귀분석에서 각 관측 값 y_i , $i = 1, \cdots, n$ 가 관측 값의 평균 $\bar{y}$ 로 부터 떨어진 정도 $y_i - \bar{y}$ 를 다음과 같이 나타낸다.

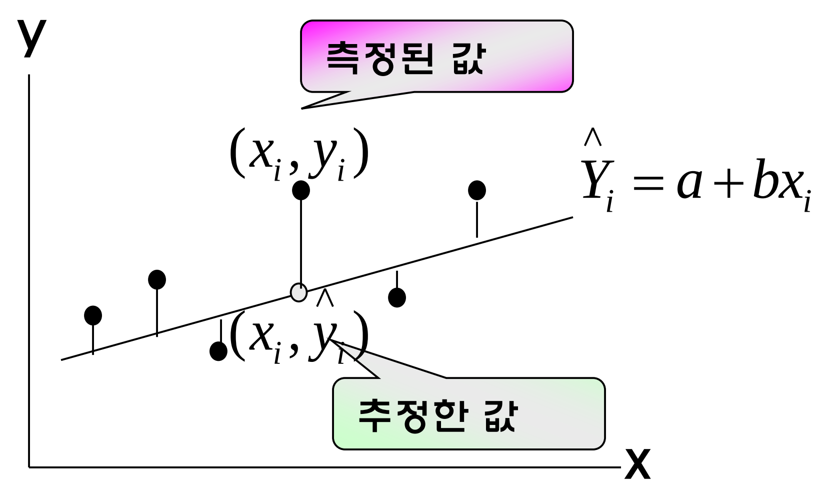
$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$



최소제곱법(Least Squares Method)에 의한 모수 추정

– e_i(잔차,Residual)의 제곱의 합을 최소로 하는 직선을 찾는다.

$$SSE=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-a-bx_i)^2$$



a와 b에 대해 SSE를 편 미분

$$\frac{\partial(SSE)}{\partial a}=-2\sum_{i=1}^{n}(y_i-a-bx_i)\equiv 0$$

$$\frac{\partial(SSE)}{\partial b}=-2\sum_{i=1}^{n}(y_i-a-bx_i)x_i\equiv 0$$

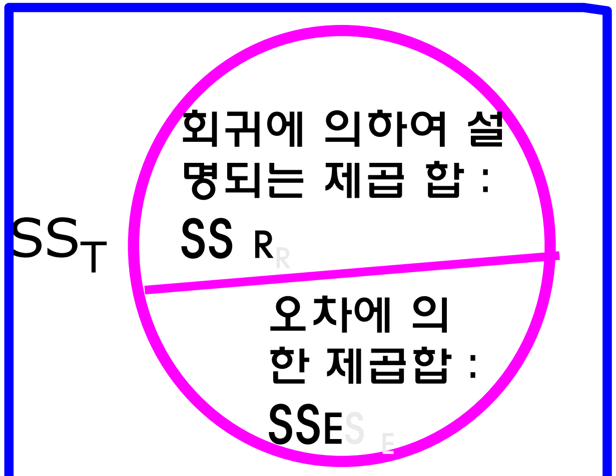
연립방정식을 a와 b에 대해 정리

$$b=\frac{n\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

$$a=\frac{\sum_{i=1}^{n}y_i-b\sum_{i=1}^{n}x_i}{n}$$

$$\boxed{\therefore \hat{Y}_i=a+bx_i}$$

앞의 식의 양변을 제곱하여 합한 뒤 정리하면 다음과 같다.



< 특성치(Y)의 총 제곱합 분해 >

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- 총 제곱 합 (SS_T)
- 잔 차(오차) 제곱 합(SS_E)
- 회귀 제곱 합(SS_R)

총 제곱 합 가운데 회귀선에 의한 제곱 합(SSR)이 차지하는 비율 R^2 을 회귀 직선의 기여율 또는 결정계수 또는 R^2 값이라고 부른다.
또한 정도를 높게 판단하기 위해서 회귀변동에서 오차분산을 뺀 순수한 회귀변동 $R^2_{(adj)}$ 를 사용하기도 한다.

$R^2 = SS_R / SS_T$; $R^2_{(adj)} = (SS_R - MSE) / SS_T$ 또는 $R^2_{(adj)} = 1 - [(SS_E/df_E)/(SS_T/df_T)]$

4.2. 회귀분석

- 자유도(¢ 또는 df)는 다른 것으로 설명될 수 없는 독립된 데이터 제곱의 갯수이고 제약조건이 있으면 제약조건의 수 만큼 자유도는 감소한다.

식	자유도	설명
$\sum_{i=1}^{n} y_i^2$	n	독립된 제곱항의 수가 n개
$\sum_{i=1}^{n} (y_i - \mu)^2$	n	독립된 제곱항의 수가 n개 제약조건이 존재하지 않음
$\sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$	제곱항의 수는 n개 $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 0$ 제약조건 존재

$$H_0 : \beta = 0 \quad H_1 : \beta \neq 0$$

• 일반적으로 회귀직선에 대한 유의성 검정은 분산분석(ANOVA) 을 이용

요인	제곱 합	자유도	평균제곱	F 값	p-value
회귀	SSR	1	MSR=SSR/1	MSR/MSE	p{F≥f}
잔차	SSE	n−2	MSE=SSE/(n−2)		
계	SST	n−1			

p-value 가 유의수준 α 보다 크면 H_0 를 기각 못함.

기각 역을 이용 시 F 값이 $F_{1-\alpha}(\phi_R, \phi_E)$ 보다 크면 H_0 를 기각,

여기서 f 는 F의 관측 값

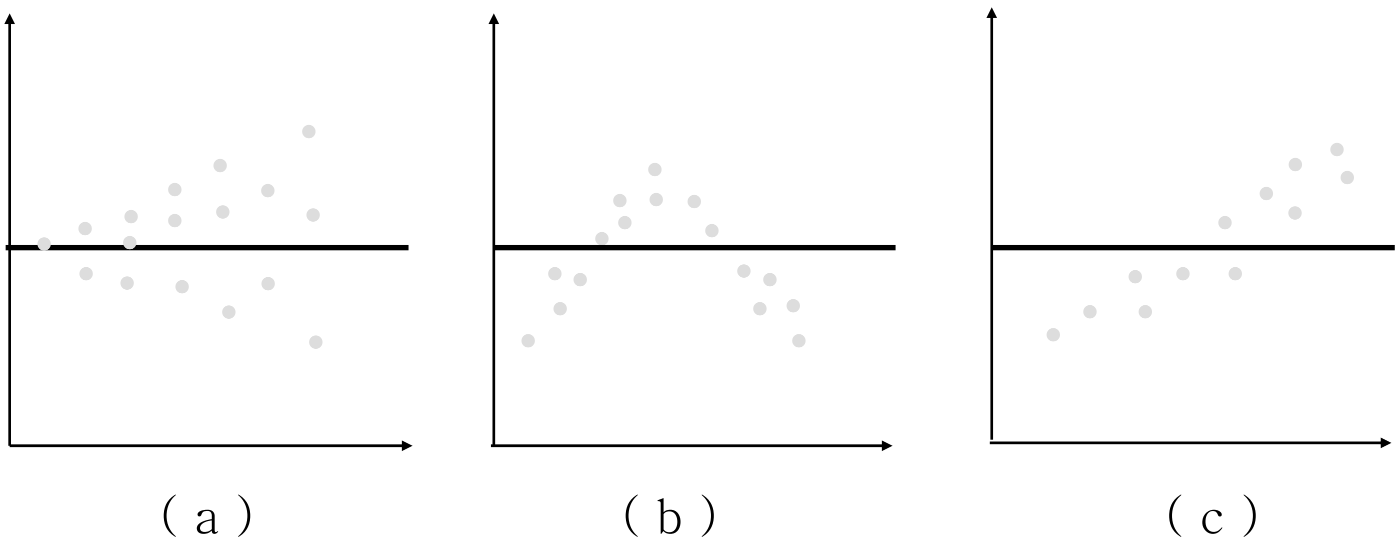
3) 잔차(Residual)의 검토

– 가정에서 벗어난 잔차의 형태

(a) 등분산성에 의심이 가는 경우

(b) 독립성 및 선형성에 의심이 가는 경우

(c) 고려중인 변수 이외의 다른 변수가 필요한 경우



4) 변수변환에 의한 회귀모형 적합

직선 방정식이 적합도(R square 등)가 나쁜 경우에 다음과 같은 변수변환을 통하여 더 좋은 방정식을 만들 수 있다.

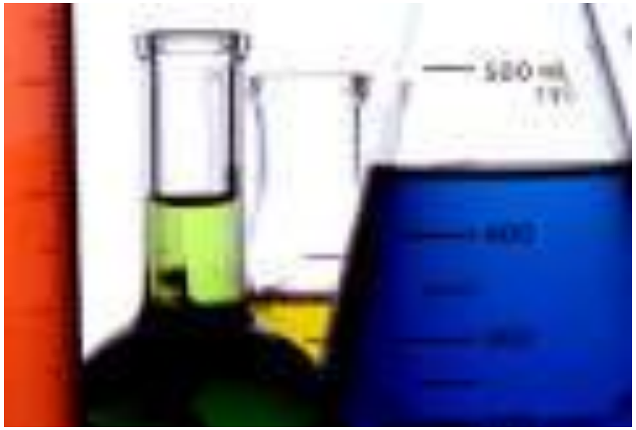
$\log Y = a + bx$	$Y = a + b\log x$	$\sqrt{Y} = a + b\log x$
$\sqrt{Y} = a + bx$	$Y = a + b\sqrt{x}$	$e^{Y} = a + \frac{b}{x}$
$\frac{1}{Y} = a + bx$	$Y = a + b\frac{1}{x}$	
$3^{Y} = a + bx$	$Y = a + b5^{x}$	

4.3. 단순회귀분석 예제

설명(독립)변수(X)가 1개이며, 반응(종속)변수 (Y)의 관계가 직선일 때

예 촉진제의 양에 따른 도금두께(반응량)의 관계를 알고자 아래의 데이터를 수집하였다. (단순회귀예제.sav)

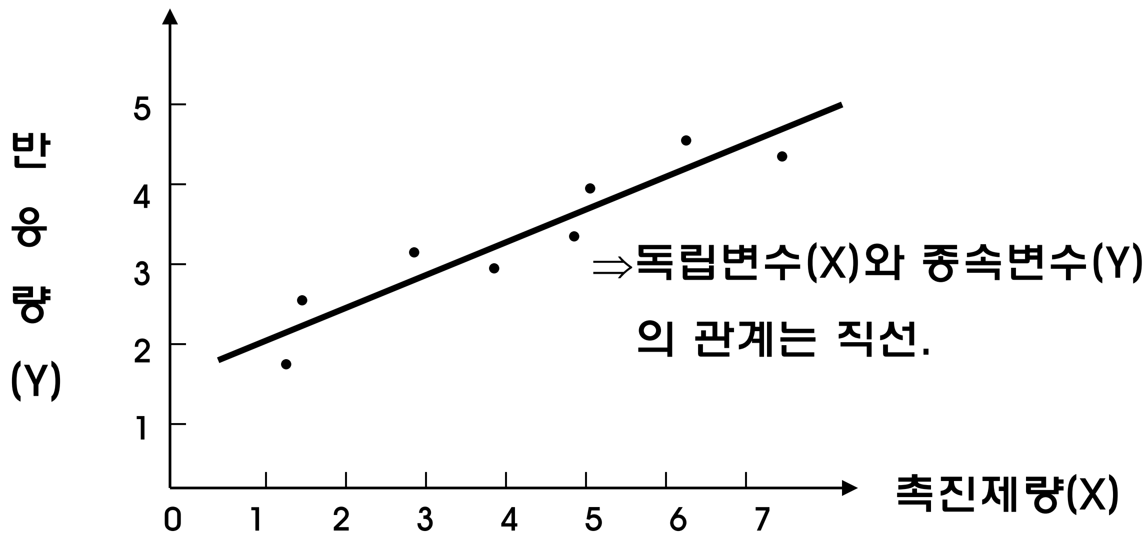
실험 번호	1	2	3	4	5	6	7	8	9	10
촉진제량(X)	1	1	2	3	4	4	5	6	6	7
반응량(Y)	2.1	2.5	3.1	3.0	3.8	3.2	4.3	3.9	4.4	4.8



4.3. 단순회귀분석 예제

단순회귀는 독립변수와 종속변수의 관계가 직선일 때, 즉 독립변수가 증가함에 따라 종속변수가 일정하게 증가하거나 감소할 때 사용한다.

(1) 산점도

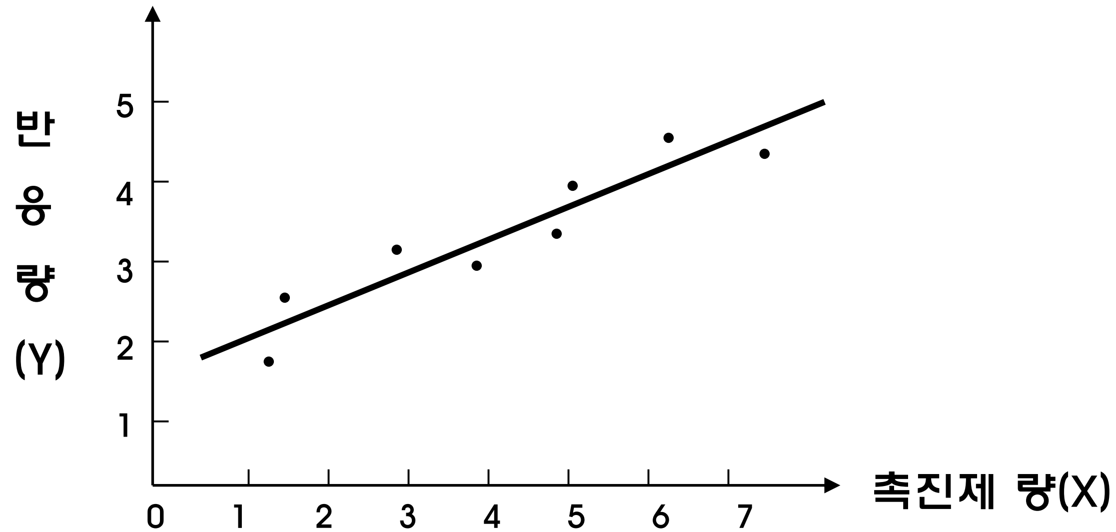


(2) 직선회귀모형 설정

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (i = 1, 2, \ldots, 10)$$

β_0 : 절편 , β_1 : 기울기 , X_i : i 번째 관측된 X값 , ε_i : 오차

(3) 방정식(회귀식) 적합



참고 최소 제곱법 (Least Squares Method)

실제 Y_i 값과 방정식에서 $\hat{Y}_i$ 의 차이의 제곱을 최소로 하도록 방정식을 추정하는 방법

Y_i 는 실제 관측된 값, $\hat{Y}_i$ 은 추정된 관계식에 의한 i 번째 Y(종속 변수) 값으로

$\sum (Y_i - \hat{Y}_i)^2$ 의 값이 최소가 되도록 관계식을 추정하는 방법

4.3. 단순회귀분석 예제

* **최소 제곱법에 의해 추정된 방정식** $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$Y_i = 2.00 + 0.387 X_i$ 혹은 $\hat{Y} = 2.00 + 0.387 X$

β_0 , β_1 : 최소제곱 법으로 추정된 회귀계수 값.

Y_i : 최소제곱 법으로 추정된 방정식에서 i 번째 Y값.

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = 0.387$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 2.00$$

추정된 방정식이 데이터에 잘 적합 되었는지를 판정하여 ,

실무에 적용할 것인지를 결정. 방법은 분산분석표에 의한 F 검정

☞ 앞의 과정을 표로 작성하여 보자.

단순회귀의 분산분석표 (유의수준 : α = 0.05)

요인	제곱 합(SS)	자유도(¢)	평균제곱(MS)	F_0	$F(1-\alpha)$
회귀	SSR (6.11)	$¢_R$ (1)	MSR (6.11)	MSR / MSE (66.284)	$F_{0.95}(1,8)$ = 5.32 P=0.000
오차	SSE (0.74)	$¢_E$ (8)	MSE (0.0925)		
계	SST (6.85)	$¢_T$ (9)			

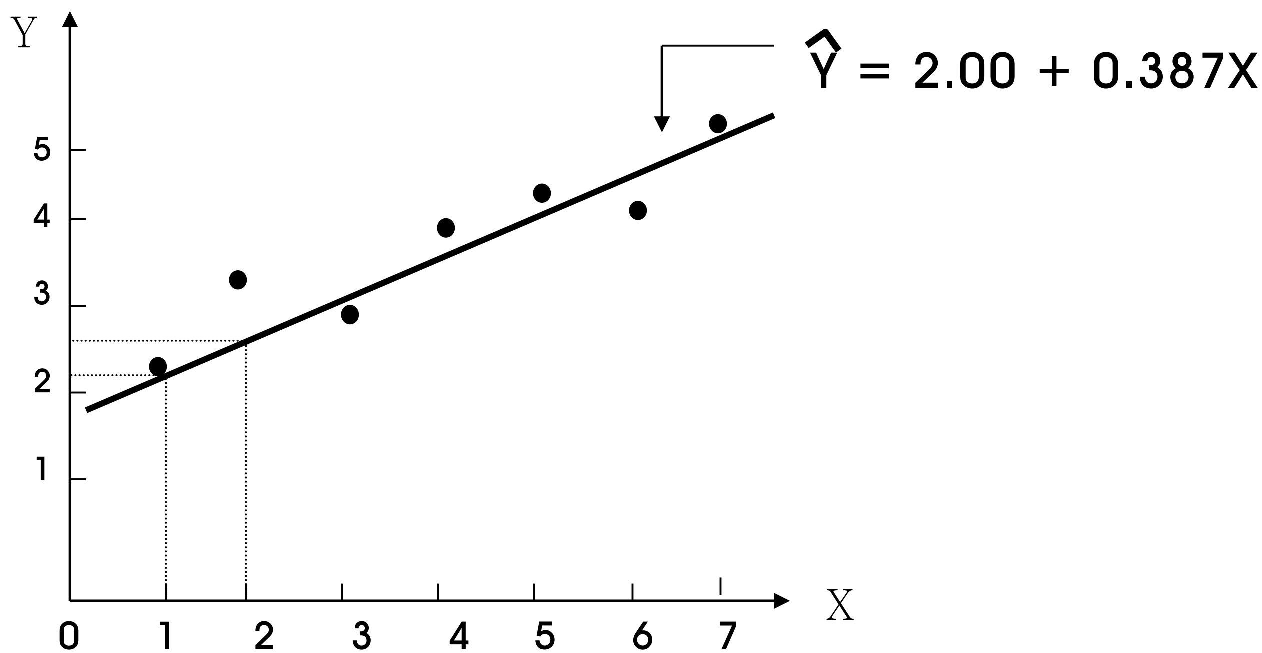
결정 계수 (R^2) $= \dfrac{SSR}{SST} = \dfrac{6.11}{6.85} = 0.89$

R−sq(수정) $= \dfrac{SSR - MSE}{SST} = \dfrac{6.11 - 0.0925}{6.85} = 0.878$

☞ 총 제곱 합 (SS_T)중에서 87.8%가 직선 회귀식으로 설명되고 있다.

회귀 방정식

F_0 가 66.05이고 F (1,8 ; 0.05) = 5.32 보다 크므로, 신뢰수준 95% (유의 수준 0.05)로 직선 회귀식 $\hat{Y}$ = 2.00 + 0.387X 는 촉진제량 (X)과 반응량 (Y)의 관계를 잘 나타내며, 촉진제량은 반응량에 유의한 영향을 준다.



▷ **추정된 $\hat{Y}$ 값.**

X	1	2	3	4	5	6	7
$\hat{y}$	2.387	2.774	3.161	3.548	3.935	4.322	4.709

$\hat{Y}_i$ (촉진제량(X)이 1일 때 반응량의 추정 값)

$= 2.00 + 0.387 X_1$

$= 2.387$

실제 관측된 반응량 2.1과 약간의 차이가 생긴다.

❑ 수정된 결정계수

– 분석 결과 해석

- 촉진제 양(X)에 따른 반응량 (Y) 사이의 추정 회귀식은 Y = 2.00 + 0.387 X 이다.

- 또한 두 변수 사이의 상관계수는 0.938로 상관관계가 매우 강하다. 단순회귀에서는 상관계수의 제곱이 결정계수 R−sq이다. 즉, 총변동 중에서 상기 회귀직선으로 설명될 수 있는 변동량이 R−Sq = 89.2% 이고, 오차를 뺀 순수한 회귀 변동량은 수정된 결정계수 R−Sq(adj) = 87.9 % 이다. 분산분석표를 이용하여 계산해 보면,

$$R\text{-}sq_{(adj)} = \frac{SS_R - MSE}{SS_T} = \frac{6.1114 - 0.0922}{6.8490} \times 100 = 87.9\%$$

단순회귀분석 실습

- 10명의 입시생들의 3월 수리영역 수능모의고사점수와 11월 수리영역 수학능력시험점수가 다음과 같다고 할 때, 3월 모의고사점수로부터 11월 수능점수를 예측하고자 한다. 어떤 분석이 적절할 것으로 보이는가?
- (수능시험.sav)

모의고사 점수	75	82	80	88	42	48	40	62	98	**44**
11월 수능점수	78	91	96	99	65	69	58	68	100	**63**

단순회귀분석 실습

- 진통제의 투여량에 따라 진통지속시간이 어떻게 변하는지 알아보기 위해 진통제의 여러 수준에서 실험한 결과가 다음과 같다.
- (진통지속시간.sav)

투여량 (DOSE)	2	2	4	4	8	8	16	16	32	32
진통지속 시간(HR)	60	58	63	62	67	65	70	70	74	73

중회귀 분석 (Multiple Regression Analysis)이란?

설명(독립)변수의 수가 두 개 이상인 경우에 반응(종속)변수와의 관계가 선형함수로 작성된 모델에 대한 분석.

주로 사용되는 회귀모형

단순회귀	설명(독립)변수 1개와 반응(종속)변수의 관계가 직선
곡선회귀	설명(독립)변수 1개와 반응(종속)변수의 관계가 곡선
중회귀	설명(독립)변수 2개 이상과 반응(종속)변수의 관계
변수선택에 의한 중회귀	설명(독립)변수가 많을 때, 중요한 변수만 찾아 회귀방정식을 적합 시킴

1) 독립변수가 2개인 중회귀모형

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

β_0, β_1, β_2 : 회귀계수, X_{1i} : X_1변수의 i 번째 관측된 값

ε_i : 오차, X_{2i} : X_2변수의 i 번째 관측된 값

2) 방정식(회귀식)적합 : 오차제곱합을 최소로 하는 회귀계수를 구한다.

$$\text{오차 제곱합} = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} + \beta_2 x_{2i})^2$$

독립변수가 2개 이상 (X_1, X_2)이고, 종속변수(Y)와의 관계를 알고자 할 때

예 **어떤 공장에서 하루에 사용되는 원료 투입량(X1)과 공정온도(X2)와 스팀의 양(Y)이 어떤 관계에 있는가를 알아보기 위하여 과거 25일간의 데이터를 수집하였다.**

다중회귀예제.sav , [단위 : X1(톤) , X2(℃) ,Y(톤)]

X_1	X_2	Y
35.3	20	10.98
29.7	20	11.13
30.8	23	12.51
58.8	20	8.40
61.4	21	9.27
71.3	22	8.73
74.4	11	6.36
76.6	23	8.50
70.6	21	7.82
57.5	20	9.14

X_1	X_2	Y
46.4	20	8.24
28.9	21	12.19
28.1	21	11.88
39.1	19	9.57
46.8	23	10.94
48.5	20	9.58
59.3	22	10.09
70.0	22	8.11
70.0	11	6.83
74.5	23	8.88

X_1	X_2	Y
72.1	20	7.86
58.1	21	8.47
44.6	20	8.86
33.4	20	10.36
28.6	22	11.08

변수선택 방법 : 독립변수의 수가 많은 경우에 사용

- 입력 : 지정한 변수를 한꺼번에 투입
- 전진 : 기준에 따라 변수를 하나씩 투입
  (Forward selection method)
- 후진 : 모든 변수를 투입한 다음, 기준에 따라 하나씩 탈락
  (Backward elimination method)
- 단계 : 각각의 단계마다 변수들을 유의도에 따라 투입, 탈락
  (가장 일반적 : Stepwise Regression method)

예 수율에 영향을 줄 수 있는 독립 변수들 가운데, 중요한 변수만 골라 회귀식을 만들고자 한다.

데이터의 수집

- 독립변수 : 농도 X_1 (%), 온도 X_2 (℃), 습도 X_3 (%), 시간 X_4 (분)
  비중 X_5 (g/㎤), 촉진제량 X_6 (g), 속도 X_7 (㎧), 압력 X_8(N/㎡)
- 종속변수 : 수율(Y)

측정번호	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
15	17 . 98	472.0	2.93	3	0.7	4	5250	205.0	10.40
16	17 . 82	460.0	3.00	3	0.7	4	5424	215.0	10.40
17	17 . 42	440.0	3.23	3	0.6	4	5345	230.0	14.70
18	19 . 47	78.7	4.08	4	0.6	1	2200	66.0	32.40
19	18 .52	75.7	4.93	4	0.8	2	1615	52.0	30.40
20	19 . 90	71.1	4.22	4	0.7	1	1835	65.0	33.90
21	20 . 01	120.1	3.70	3	0.7	1	2465	97.0	21.50

4.4. 중회귀분석

선택되는 변수의 개수	선택되는 변수의 번호	F_0	C(p)	R^2
1	X_2	2.7	2.7	0.602
2	X_2, X_3	6.9	3.6	0.741
3	X_2, X_3, X_5	23.2	3.8	0.854
4	X_2, X_3, X_5, X_6	21.2	5.7	0.864
5	X_2, X_3, X_5, X_6, X_8	15.1	6.8	0.872

결론 3번째 회귀식이 가장 좋음.

: F_0 와 R^2 값이 크고 C(p) 는 (K(변수)+1)에 근접하는 값은 3번째 식으로, 변수의 개수가 3개로 적절하다.

중회귀 방정식

변수선택에 의한 중회귀식 $\hat{Y} = 61.41 - 3.2X_2 + 2.94X_3 + 2.1X_5$

4.5. 중회귀분석 예제

자동차 타이어의 실내 주행실험에 있어서 타이어에서 발생되는 열은 다음과 같은 5가지 변수에 의하여 영향을 받는 것으로 알려져 있다.

X1 : 타이어에 걸리는 하중
X2 : 속도(km/hr)
X3 : Shoulder의 두께(mm)
X4 : 실내온도
X5 : 측정시간(min)
Y : 발열량

발열량에 영향을 미치는 변수를 찾고 회귀모형을 구축해보자.
<타이어.sav>

OBS	X_1	X_2	X_3	X_4	X_5	Y
1	70	70	36.5	36	5	91
2	70	70	36.0	36	6	89
3	70	90	37.0	37	6	105
4	70	90	36.3	37	6	106
5	70	110	36.5	39	4	113
6	70	110	36.0	39	5	114
7	90	70	36.5	38	5	117
8	90	70	36.3	38	6	115
9	90	90	36.6	39	5	125
10	90	90	36.6	39	6	126
11	90	110	37.0	38	6	140
12	90	110	35.6	38	6	141
13	110	70	35.3	38	7	140
14	110	70	36.8	35	7	142
15	110	90	35.3	38	5	150
16	110	90	35.3	38	6	149
17	110	110	37.1	38	4	168
18	110	110	35.6	37	5	166

Q/A

㈜ 아이티베인
이현우
hyunwoo@itvane.co.kr , 010-5245-1653