목차

[Ch01 통계 1](#_Toc11014768)

[Ch02 표본조사 2](#_Toc11014769)

[Ch03 자료 4](#_Toc11014770)

[Ch04 기술통계분석 7](#_Toc11014771)

[ Ch04-1 집중화 경향(central tendency) 에 대한 기법 7](#_Toc11014772)

[ Ch04-2 데이터가 어떻게 퍼져 있는지를 설명하는 기법 8](#_Toc11014773)

[ Ch04-3 자료의 정리 (표와 그래프) 9](#_Toc11014774)

[Ch05 확률 10](#_Toc11014775)

[Ch06 이산 확률분포 12](#_Toc11014776)

[Ch07 연속확률분포 14](#_Toc11014777)

[Ch08 통계적 추론 15](#_Toc11014778)

[ Ch08-1 추정 15](#_Toc11014779)

[ Ch08-2 가설검정 15](#_Toc11014780)

[Ch10 통계분석방법 18](#_Toc11014781)

[Ch11 평균차이검정 25](#_Toc11014782)

[Ch12 분산분석 25](#_Toc11014783)

[Ch13 회귀분석 26](#_Toc11014784)

[Ch14 교차분석 26](#_Toc11014785)

Ch01 통계

* 통계학이란?
  + 관심대상인 **모집단의 특성**을 파악하기 위해,
  + 표본을 수집하고

🡺 **표본추출**(sampling)

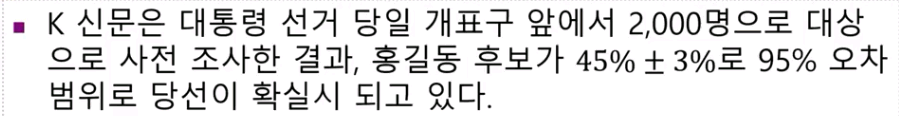
* + 수집된 자료를 **요약** 하여 표본의 특성을 파악하고

🡺 **기술통계학**(descriptive statistics)

* + 표본의 자료를 이용해서 모집단의 특성에 대해 확률(95% 오차범위)을 이용해 추론(정보화)하는 학문이다.

🡺 **추리통계학**(inferential statistics)

* 기술통계학
  + 수집된 자료를 정리, 요약하여 수치, 표, 그래프로 자료의 특징을 정리하는 과정
  + 자료가 가지고 있는 원래의 특성만 표현
  + 고등학교 1학년 평균 키. 대통령선거에서 각 후보가 받은 득표수.
* 추측통계학
  + 샘플을 추출, 분석하여 그 결과로 전체 모집단의 특성을 확률적으로 판단(추측)
  + 새로운 가설이 맞는지 틀리는지를 검증
  + 5000개 형광등에서 50개 샘플로 조사한 형광등 수명으로 불량품 예측



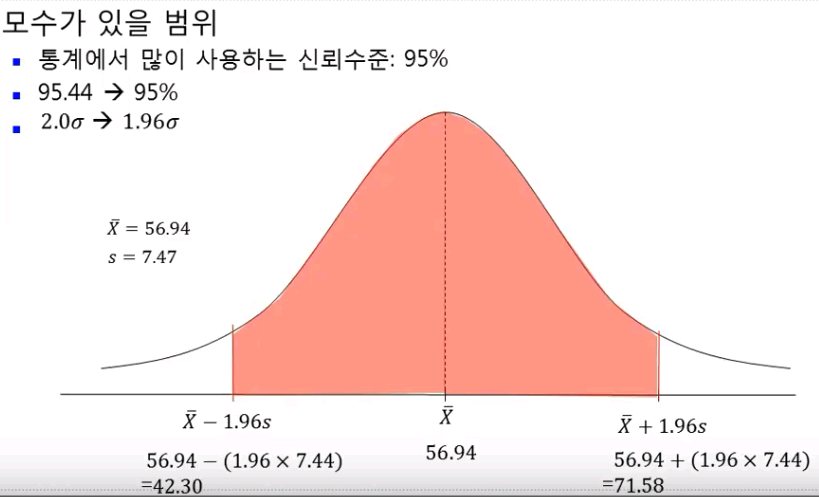


* 오차범위와 통계
  + Empirical Rule(경험법칙): 표본에서 평균과 표준편차를 구하면 확률을 통해서 오차범위를 구할 수 있다.

평균으로부터 1표준편차에 전체데이터 68%가 있다.

평균으로부터 2표준편차에 전체데이터 95.44%가 있다.

평균으로부터 3표준편차에 전체데이터 99.73%가 있다.



Ch02 표본조사

* 자료 수집방법 – 실험, 조사, 출판자료
* 자료 조사방법
  + 전수조사(census)
  + 표본조사(sample)
* 모집단(Population N)과 표본(Sample n)
  + 모집단 (Population): 연구자가 알고 싶어하는 대상 / 집단 전체
  + 표본 (Sample): 연구자가 측정 또는 관찰한 결과들의 집합

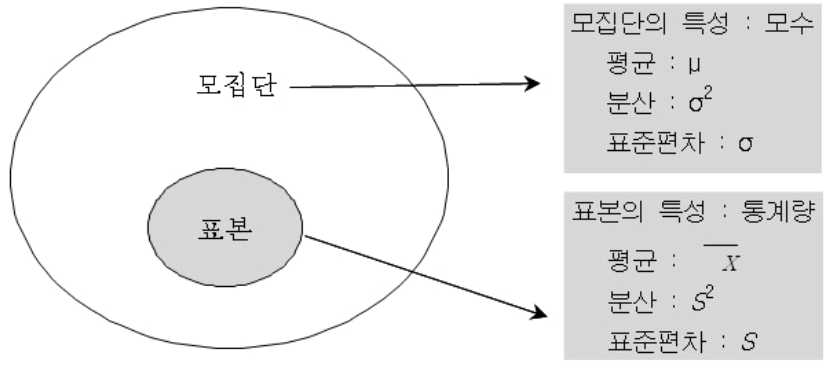
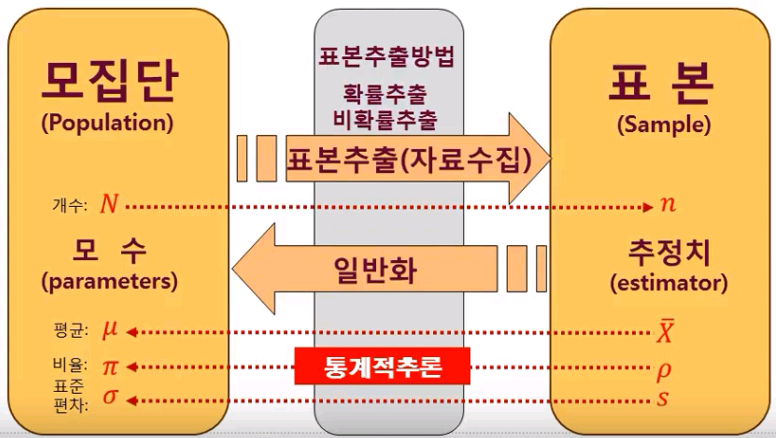
전수조사는 거의 불가능하고 매우 비효율적.

따라서 표본 (Sample)을 측정 또는 관찰해서 알고 싶어하는 모집단(효과/대상)을 추정.

* 모수(parameter)와 통계랑(statistics)
  + 모수

- 모집단이 가지고 있는 특성의 실제 값. 평균신장, 평균수입

* + 통계량(statistics)=추정치(estimates)=통계치
    - 표본에서 얻은 수치 특성 값. 표본에 따라서 값이 달라진다 (**오차발생**)

* 모수와 통계량

|  |  |
| --- | --- |
| 모수 | 통계량 |
| 모집단에 대한 수치의 특성 값 | 표본에서 얻은 수치 특성 값 |
| 상수 = 모집단은 진실된 하나의 값 | 확률변수 = 표본에 따라 변하는 값. |
| 모 평균, 모 표준편차 | 표본 평균, 표본 표준편차 |
| 우리나라 고등학교 사교육비 평균 | 우리나라 고등학교 1학년 중에서 10000 명만 뽑아서 조사하여 얻은 평균 사교육비 |

* 표본오차 와 통계적 추론 (sampling error)
  + 모집단에서 표본을 추출해서 조사하기 때문에 모수와 표본 통계량 사이에 생기는 오차
  + 표본의 크기를 클수록 표본오차감소.
  + 표본오차의 허용범위를 확률로 구하는 것이 통계의 목적(통계적 추론)
* 통계적 추론

실제 알고자 하는 값은 모집단의 모수

전수조사가 불가해서 표본조사

실제 모수와 표본조사의 통계량과 오차가 발생

그 오차범위를 확률을 이용해서 전체 값을 예측(추론)

* 표본조사방법
  + 확률추출법

사전에 일정한 추출확률이 주어지는 표본추출법. 표본 추출틀(모집단의 조사대상 요소의 명단)이 존재

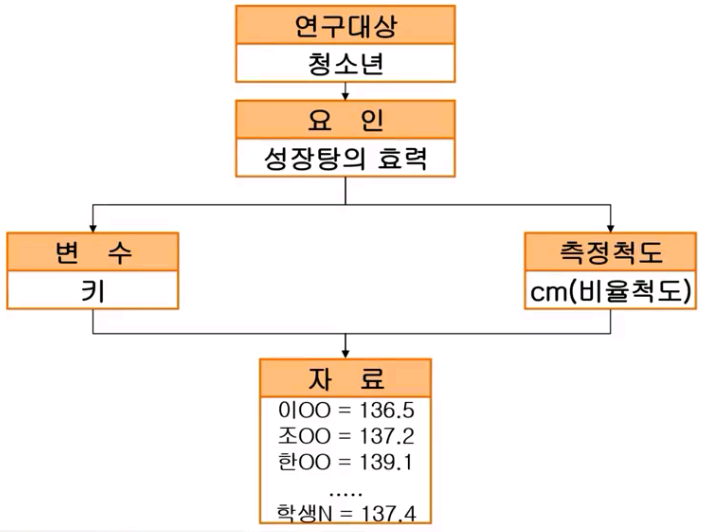
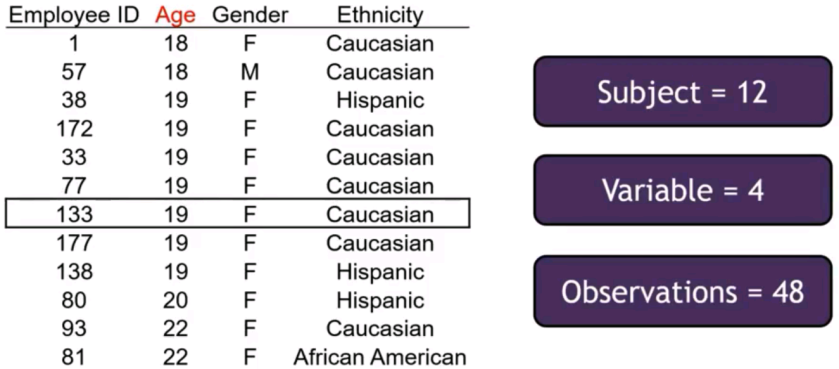
1. 단순확률추출(simple random sampling)
   * + - 표본이 뽑힐 확률이 동일한 상황에서 뽑는 방법
       - 모집단 전체에 대한 추출틀(명단)이 있어야 한다.
       - 현대카드 소지자 1,000만 명의 명단을 이용해서 10,000명을 난수로 추출.
2. 계통추출법(systematic sampling)
   * + - 모집단의 추출틀에서 X번째 간격 마다 하나씩 표본으로 추출
       - 형광등 불량품을 조사하기 위해 5,000개의 slot에서 20개를 샘플로 추출하려고 할 때
       - 레스토랑 손님만족도 조사: 10번째 손님, 20번째 손님
3. 층화추출법
   * + - 표본추출을 위해 모집단을 서로 겹치기 않게 몇 개의 부분으로 나눈 후, 단순확률추출
       - 서울시장 후보 선호도 조사를 위해 1000명을 조사할 때 강서구 인구비율이 10%면 100명 추출
4. 집락추출법
   * + - 조사 단위가 산재되어 조사 비용이 많이 들 때
       - 서울 대학생들의 평균 용돈조사. 모집단: 서울에 있는 56개 대학 대학생 전체이나 그 중에서 몇 개 학교만 무작위로 추출.
   * 비 확률추출법
     + 편의 추출
     + 할당 추출
     + 포커스그룹

Ch03 자료

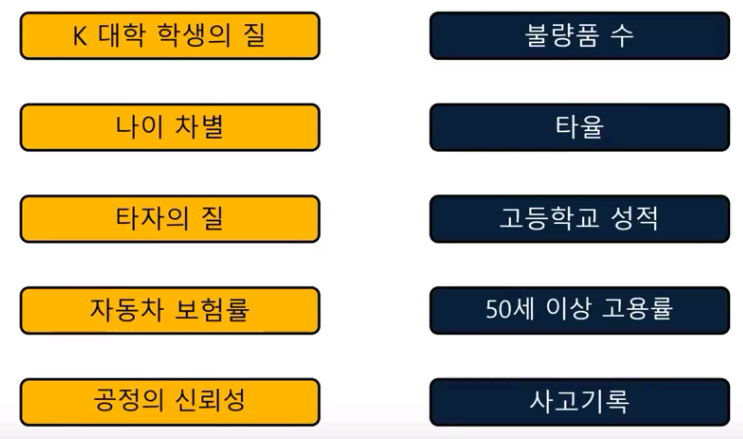
* 자료



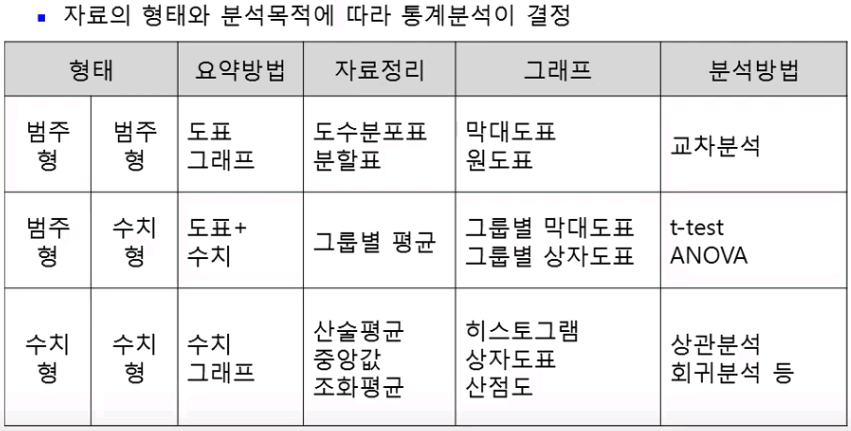
* + 개체(subject) – 연구자가 관심을 갖는 **대상**.
  + 요인(Factor) 또는 개념(concept) – 개체에 대한 특성 중 연구자가 관심을 갖는 **특성**.
  + 변수(Variable) – 요인을 구성하고 있는 요소. 변수는 일정한 특정 단위 또는 숫자로 표현이 가능해야 한다.
  + 척도(Measurement) – 측정하기 위해서 사용한 측정도구. 키-cm, 몸무게-kg으로 혈액형- A, B, O 성별-F/M

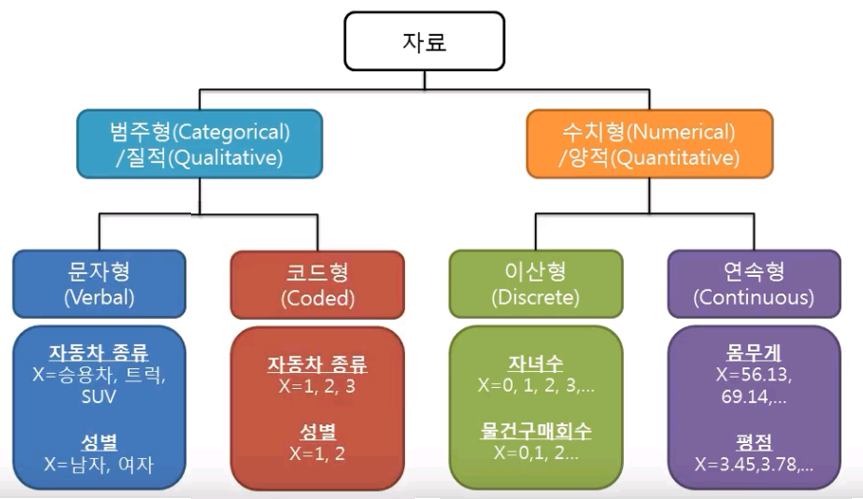
 

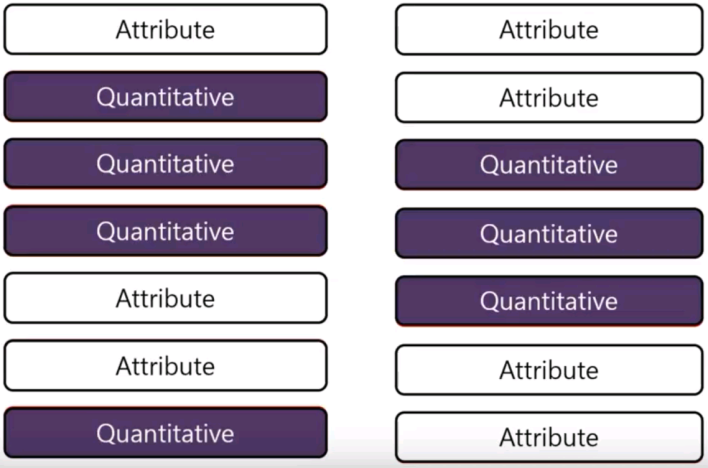
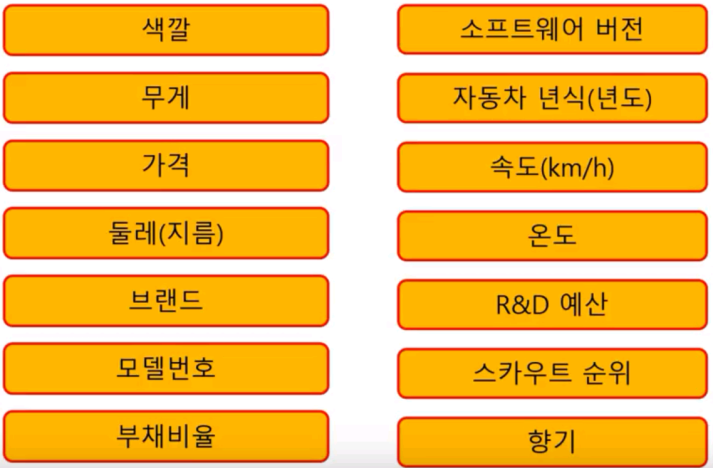
* 요인(concept) VS 변수(Variable)



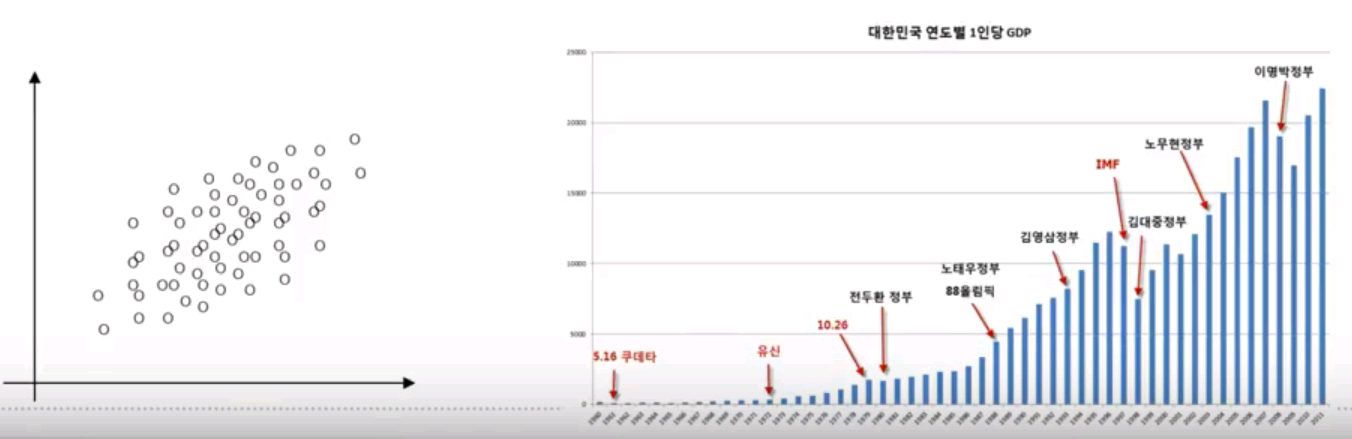
* 자료와 분석방법 - 자료에 따라 분석방법이 결정된다. 따라서 자료의 종류를 구별할 수 있어야 한다.



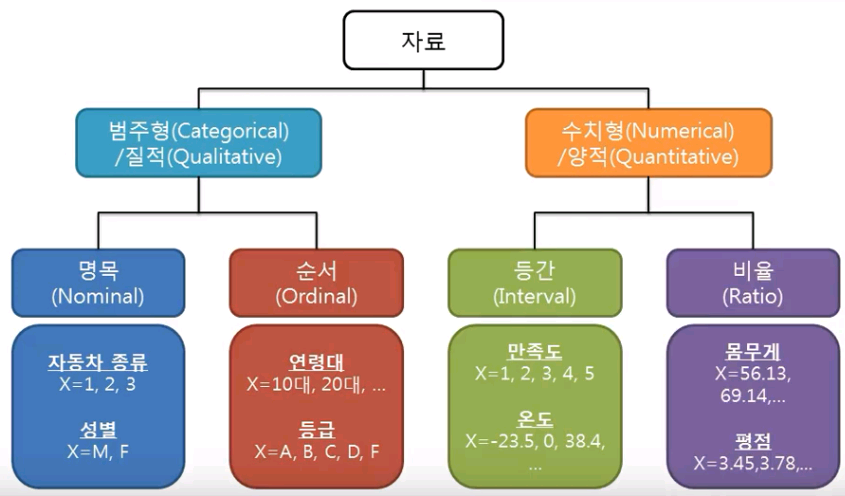
* 자료의 종류
* 
  + 질적 자료 (qualitative data, categorical data)
    - 몇 개의 특성에 따른 그룹으로 나누어 코드 숫자로 나타낸 자료
    - 성별(남자=1, 여자=2) 연산불가.
  + 양적 자료(quantitative data, Numerical data)
    - 이산형 자료 (Discrete data) - 셀 수 있는 자료 ex. 도시 의 아파트 층 수, 썩은 치아 수
    - 연속형 자료 (continuous data) 구간에서 값을 모두 취할 수 있는 자료. Ex. 세금 액수



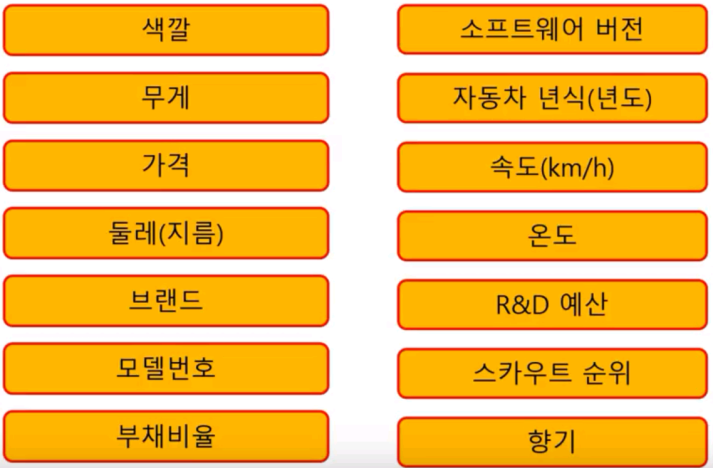
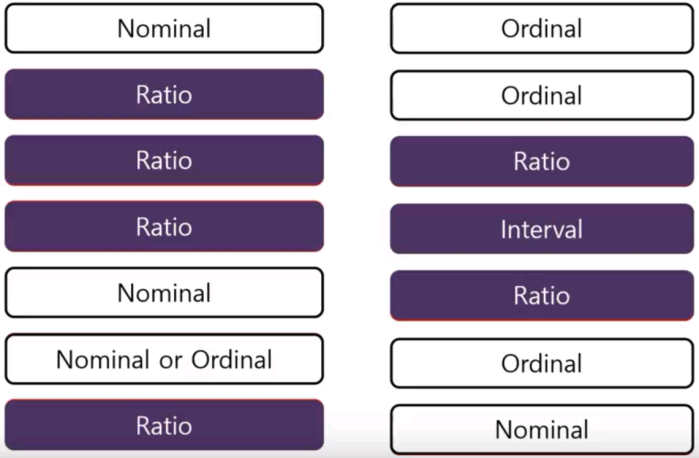
* 자료의 종류(시간 관점)
  + 종단 자료(Cross-sectional data) – 한번의 시간에 얻어진 데이터
  + 횡단 자료(Longitudinal data) – 여러 시간에 걸쳐 얻어진 데이터 Ex. 주식, 2010 ~ CDP변화



* 척도(Measurement )의 종류



* + 명목척도(**Nominal**)
    - 명목 척도는 숫자로 표현될 수 있지만 수량적인 의미를 갖지 않고, 그룹(카테고리)를 구분하는 용도로 쓰이는 척도
    - 성별(남/여) 국적(한국, 일본, 중국 등), 직업(회사원, 공무원, 자영업 등)
  + 순서척도(**Ordinal**)
    - 속성에 따라 순위를 결정하는 척도.
    - 서열자료. 순서관계가 성립하는 경우. ex. 연령대, 학년
  + 등간척도(**Interval**)
    - 절대영점이 존재하지 않지만 속성을 평가할 수 있는 균일한 간격을 두고 측정하는 척도.
    - 속성의 차이를 양적인 차이로 측정하기 위하여 척도간 간격을 균일하게 분할하여 측정하는 척도
    - Ex 설문지의 설문문항(리커르트 5점), 온도, 아이큐지수
  + 비율척도(**Ratio**)
    - 절대 영점이 있는 등간 척도. 비율자료. 양적인 차이를 나타내 주는 변수. Ex. 시험점수, 키

Ch04 기술통계분석

기술 통계에서 "기술"은 영어로 Descriptive 입니다. 한국말로는 "묘사하는" 또는 "그려서 설명하는"이라는 뜻입니다. 즉, 기술 통계는 우리가 수집한 데이터를 묘사하고 설명하는 통계 기법들을 말합니다.

그렇다면 기술 통계에는 어떤 것들이 있을까요? 다시 말해 **수집한 데이터를 설명하는 통계 기법**이 무엇이 있을까요? 기술 통계 기법은 크게 또 두 가지로 구분할 수 있습니다. 하나는 우리가 수집한 데이터를 대표하는 값이 무엇인지 또는 어떤 값에 집중되어 있는지를 다루는 기법입니다. 조금 전문적으로 말씀 드리면 데이터의 **집중화 경향 (Central tendency)에 대한 기법**이라고 말할 수 있습니다. **평균 (mean), 중앙값(median), 최빈값(mode)** 등이 바로 집중화 경향에 속하는 것들 입니다.

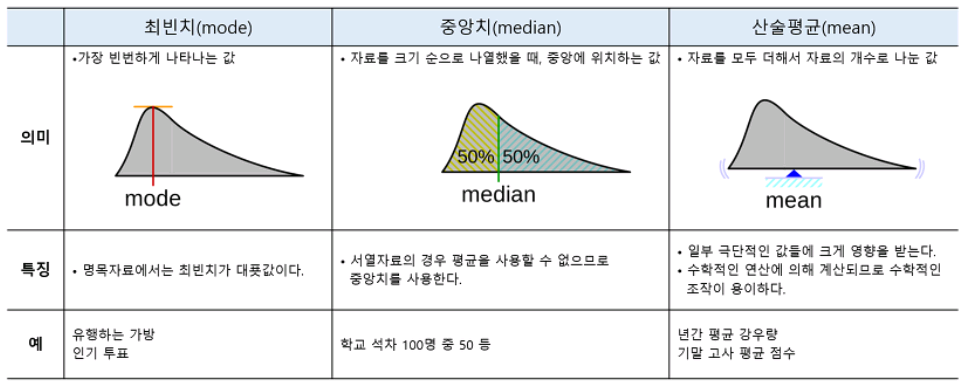
다른 하나는 우리가 수집한 **데이터가 어떻게 퍼져 있는지를 설명하는 기법**이 있습니다. 이를 분산도 (Variation)라고 부릅니다. 분산도는 말 그대로 데이터가 전반적으로 어떻게 분포되어 있는지 즉, 뭉쳐있는지 퍼져 있는지를 설명하는 방법입니다. 대표적으로 **표준편차 (standard deviation), 사분위(quartile)** 값 등이 있습니다.

우리는 기술 통계 기법을 통해서 수집한 데이터의 전체적인 모양을 그릴 수 있습니다. 예컨대, 우리나라의 국민 1인당 평균 소득이 2만 달러라고 합시다. 이 값은 우리나라 국민의 소득 수준의 대표값입니다. 하지만 대표값 만이 중요할까요? 아닙니다. 분산도 중요합니다. 예를 들어 국민 1인당 평균 소득이 아무리 높아도, 소득에 대한 편차도 함께 높은 값을 가지고 있다고 가정해 봅시다. 편차가 크다는 것은 소득 분포가 넓게 분포되어 있다는 뜻이고, 이는 다시 말해 국민의 소득 편차가 크다는 뜻입니다. 이를 통해 "소득의 분배가 잘 이루어지지 않고 있구나" 하고 해석할 수도 있습니다.

* Ch04-1 집중화 경향(central tendency) 에 대한 기법

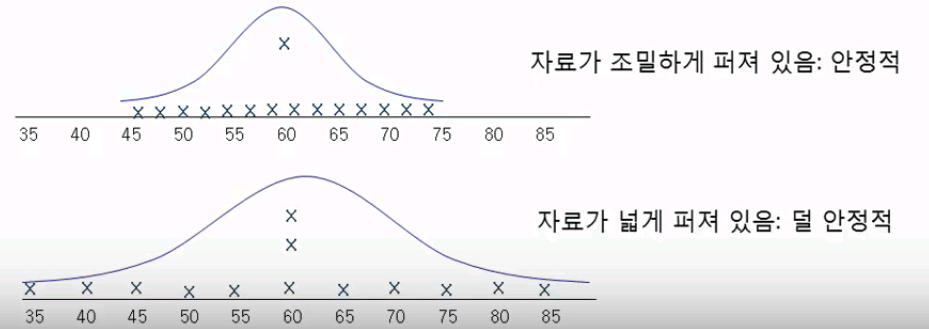
수집한 자료 전체를 대표하는 값이 무엇인지 나타내는 통계.

* + 중앙값: 자료를 크기 순으로 정렬했을 때, 중앙에 위치하는 값 (순위 자료의 대표값)
  + 최빈값: 수집한 데이터 중 그 빈도가 가장 많이 나타나는 데이터 (명목 자료의 대표값)
  + 산술평균값: 자료를 모두 더해서 전체 자료의 개수로 나눈 값 (정량적 자료의 대표값)

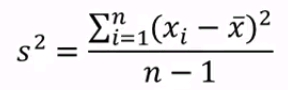


* + 가중평균: 각 항의 수치에 그 중요도에 비례하는 계수를 곱한 다음 산출
  + 기하평균: 곱의 형태로 변화하는 자료 ex.물가상승률, 인구변동률, 연평균 증가율
  + 산술평균의 문제
    - 평균값은 이상 치에 의해 민감하게 영향을 받는다. 실제로 자료가 어떻게 변화하는지 알 수 없다.
    - 중앙값, 최빈값, 절단평균으로 보완
* Ch04-2 데이터가 어떻게 퍼져 있는지를 설명하는 기법

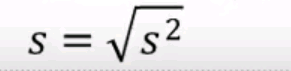
자료가 중심위치로부터 어느 정도 흩어졌는지, 중심위치(평균)이 얼마나 안정적인지에 대한 정보

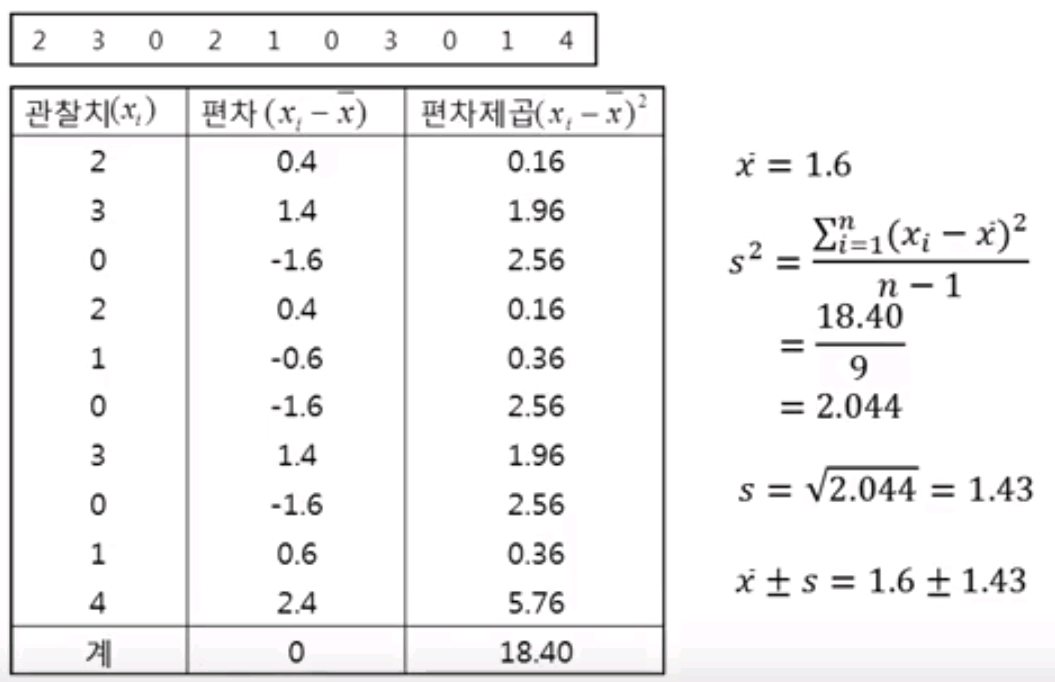


* 분산(variance)
  + - 자료가 평균을 중심으로 얼마나 광범위하게 분포하고 있는 가를 하나의 수치로 나타낸 통계량
    - 평균은 모든 데이터의 중심이니까 거리를 다 더하면 0이 되기 때문에 +, - 를 없애려고 제곱해서 루트

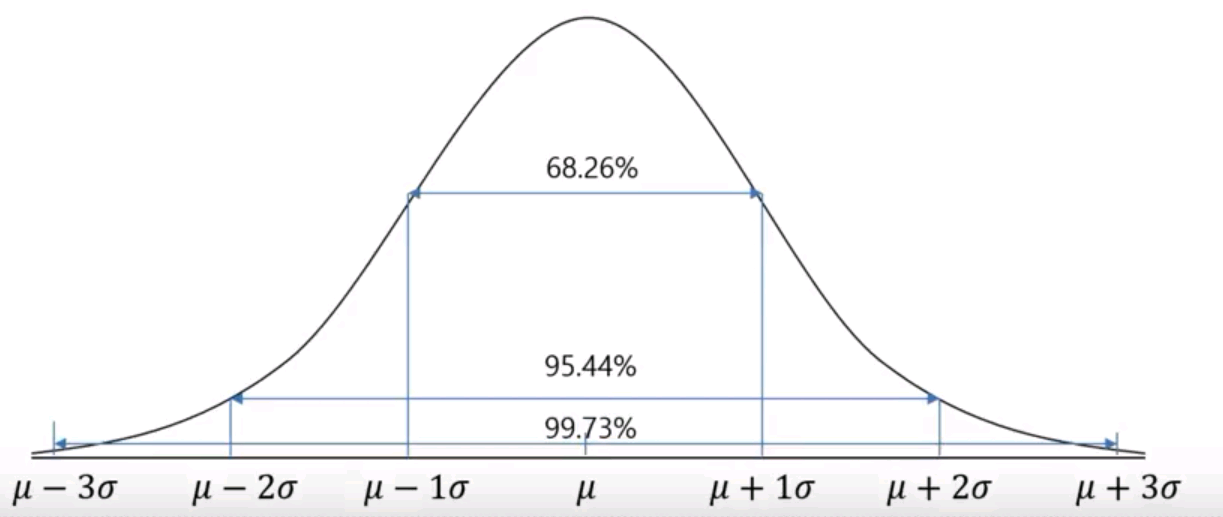


* 표준편차(standard Deviation) – 평균으로부터 자료가 흩어진 정도





* 표준편차의 중요성
  + 자료의 분포와 변동에 대한 중요한 정보를 제공
  + 통계학의 중요한 규칙과 연결됨. Empirical Rule (경험적 법칙)



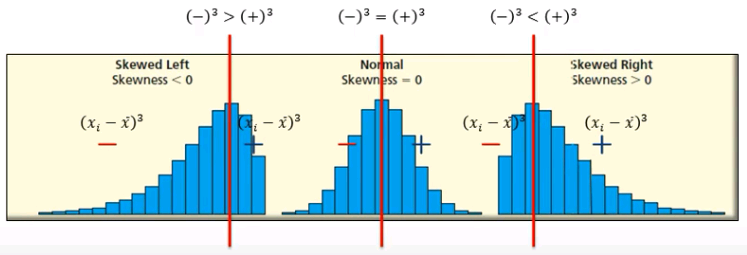
* + - 68.26% 이상의 데이터가 사이에 있음
    - 95.44% 이상의 데이터가 사이에 있음
    - 99.73% 이상의 데이터가 사이에 있음
* 분포의 형태로 자료의 이상 점이 있는지 점검

변인의 분포가 정상분포곡선으로부터 심하게 벗어난 것은 통계의 '정상성(normality)' 전제를 어기는 것. 정상분포의 전제를 어길 경우 통계 값을 신뢰하기 어렵고, 표본의 결과가 모집단의 결과인지를 추리하기 어렵기 때문에 모수통계방법을 사용하지 않는 것이 바람직.

* + 왜도

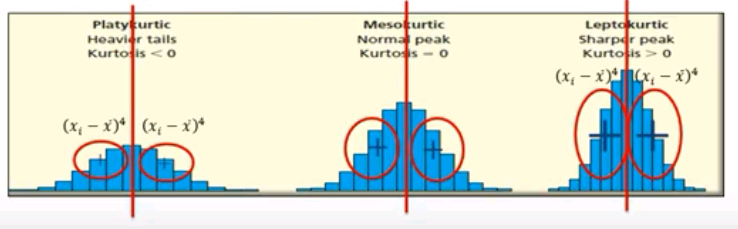
데이터가 대칭이 아닌 정도. 값(0, 양수 또는 음수)이 데이터 형상에 대한 정보를 나타냄.

꼬리가 오른쪽으로 길게 뻗고 있고 두툼한(빈도가 높은) 부분이 왼쪽이라면 낮은 값에 빈도가 높다는 뜻. 왜도가 0보다 크다면 변인의 분포가 정상분포곡선보다 왼쪽으로 치우진 것. 소득비율

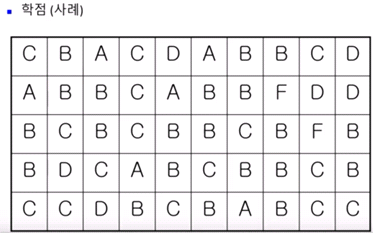


* + 첨도

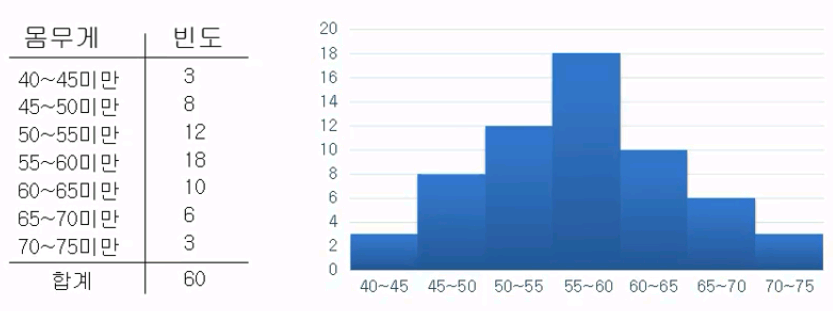
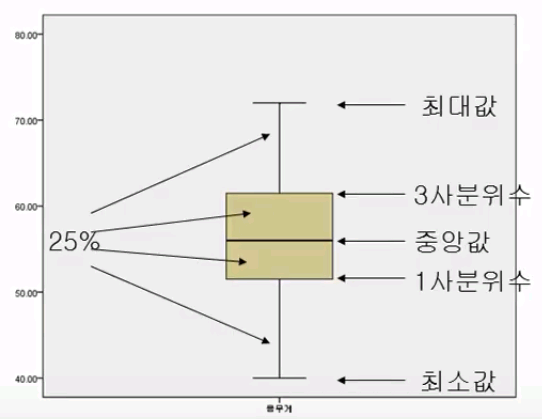
첨도는 분포의 정점과 꼬리가 정규 분포와 어떻게 다른지 나타냄. 값(0, 양수 또는 음수) 0보다 크다면 상당수가 평균값 근처에 몰려있다는 뜻



* Ch04-3 자료의 정리 (표와 그래프)
* 도수분포표(frequency table) & 막대그래프
  + 수집된 자료를 적절한 계급으로 분류, 정리한 표
  + 데이터 각 값의 출연도수를 세거나 몇 개의 구간으로 나누어 각 구간에 속하는 데이터의 개수를 세어서 정리한 표

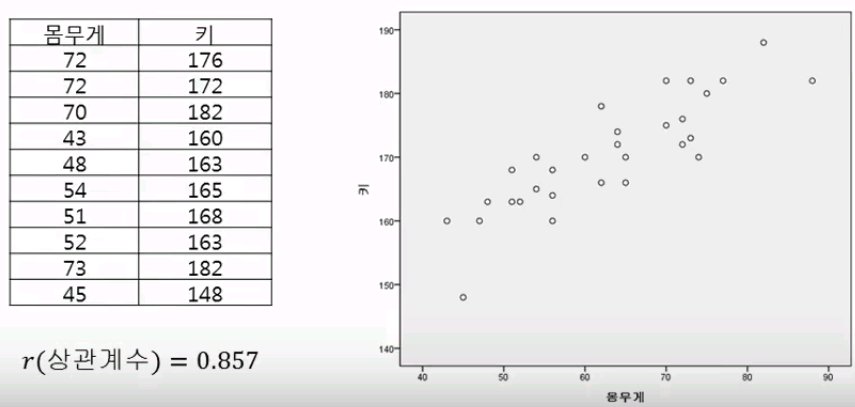
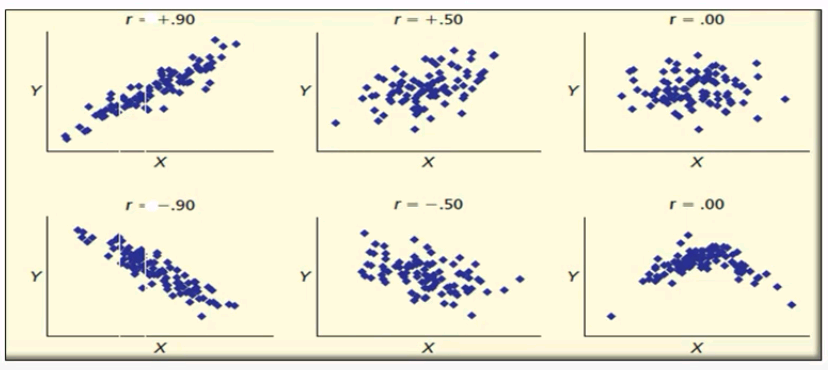
* 수치형 자료 (히스토그램 & 상자도표 )

* 상관관계 자료(산점도)

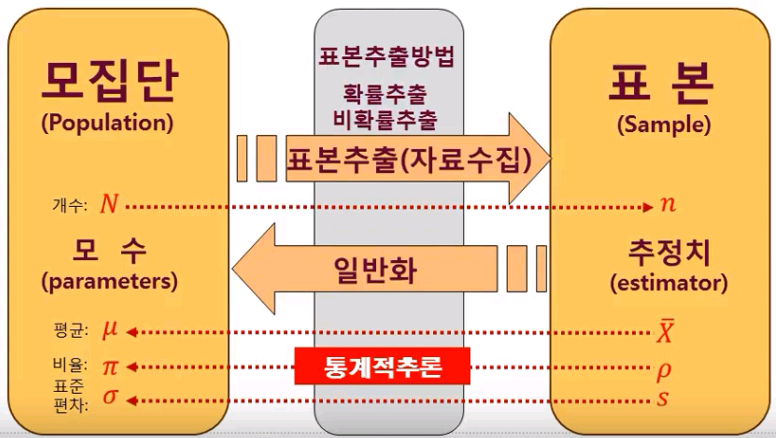
두 변수간 상호의존 관계가 있을 때, 하나의 변수에 대한 정보를 가지고 다른 변수를 예측하거나 설명.

키와 몸무게의 관계, 야구선수의 홈런과 연봉액수, 라면의 선전비용과 판매량, GNP와 자동차 보유대수

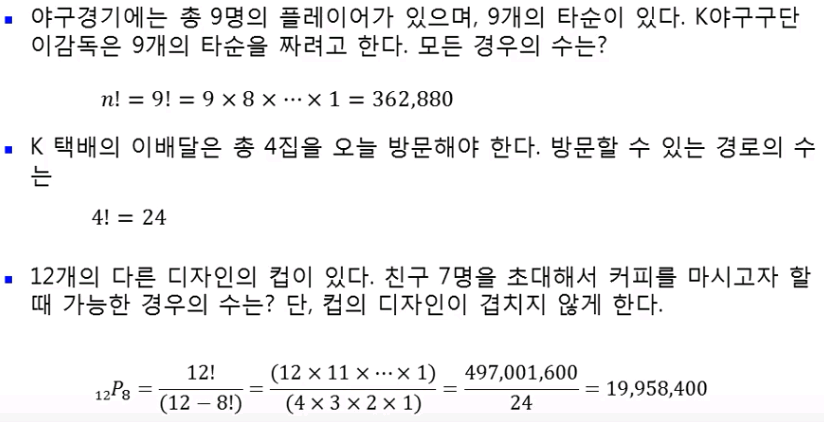
 

Ch05 확률

* 통계학이란? 표본의 자료를 이용하여 모집단의 특성에 대해 확률을 이용해 추론



* 확률
  + 통계적 확률
    - 표본 수에 따라 비율이 뀌지만 최종적으로 모수에 수렴. 즉, 표본이 무한히 커지면 모집단으로 수렴
    - 확률은 표본의 특성을 구하는 것이 아니라 모집단의 특성을 구하는 것
    - 즉, 확률은 모집단의 특성을 찾는 것,
  + 확률을 이용해서 모수를 측정
    - 추정 – 신뢰수준(95%)
    - 가설검증 – 유의수준(5%)
* 카운팅
  + 확률 = 가능성 = 경우의 수 = 카운팅
  + 곱의 법칙 (경우의 수)
    - 주사위 2개를 던질 때 나오는 경우의 수? 6\*6=36
    - 부사장3, 이사4, 부장5가 있는 회사에서 대표 1명씩 총 3명의 위원회가 열릴 경우의 수? 3\*4\*5=60
  + 순열: 서로 다른 n개 중 r개를 선택해 순서를 고려해 나열하는 방법의 수



* + 조합: 서로 다른 n개 중 r개를 선택해 순서를 고려하지 않고 뽑는 방법의 수 (순열에서 중복 제거)
    - 45개의 번호를 가진 로또에서 6개의 숫자를 고를 경우의 수는?
      * 45C6 = 45P6/6! = 5,864,443,200/720=8,145,060
      * 1등 당첨확률 = 1/8145060 = 0.00000012
      * 이중 6개 중에서 3개의 숫자가 동일할 경우 5등이다. 경우의 수와 확률은?
        + 6C3 X 39C3

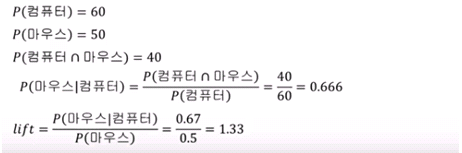
= ((6\*5\*4)/(3\*2)) \* ((39\*38\*37)/(3\*2)) =182,780 45개중 6개의 숫자를 선택

* + - * + 182,780 / 8,145,060 = 0.022
* 확률 규칙
  + 확률 - 전체 중에서 사건이 차지하는 크기. 1을 넘을 수 없음
  + 비율(proportion) 전체 중에서 사건이 차지하는 크기
  + 비(ratio) 사건 A와 B의 상대적 크기. 성비, 인구밀도. 1을 넘을 수가 있음
  + 율(rate) 특정 시간도안 발생한 사건과 노출된 횟수 (1/1,000) \* 비율과 같은 의미로 많이 사용
    - 총인구52,341,015, 여자 25,651,015, 남자 25,663,668
    - 여자비율=25,651,015/52,341,015=49.99
    - 여성비= 25,651,015/25,663,668=99.56
  + 승산비(odd ratio) 사건이 발생할 확률과 발생하지 않을 확률간의 비율
    - 경주마에 대한 승률이 4:1이라고 했을 경우 승산비와 승률은?
      * 승산비 4/1 = 4
      * 승률 1/5= 0.2
  + **조건부확율(conditional probability)** 
    - A라는 조건이 주어진 상태에서 B가 발생할 확률.
    - 시간의 관점에서 A라는 사건이 이루어진 이후에 B라는 사건이 발생할 확률
    - 교차분석, 데이터 마이닝의 연관성분석 중요한 개념

|  |  |
| --- | --- |
| A라는 조건이 추가되었을 때 B가 발생 할 확률 (표본공간이 바뀜). |  |
| P(A) = 60/120  P(B|A) = 10/60 |

* + - 매장에 방문한 모든 사람에게 마우스를 사라고 추천한다면? P ( B )

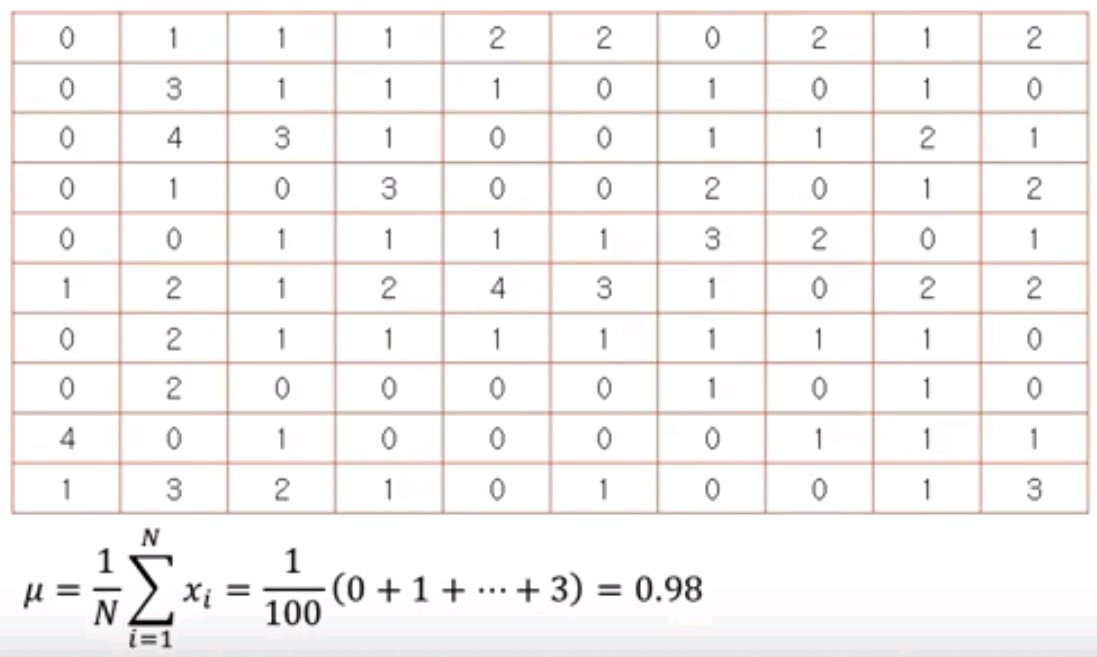
컴퓨터를 산 사람에게 마우스를 사라고 추천한다면? P ( B|A )

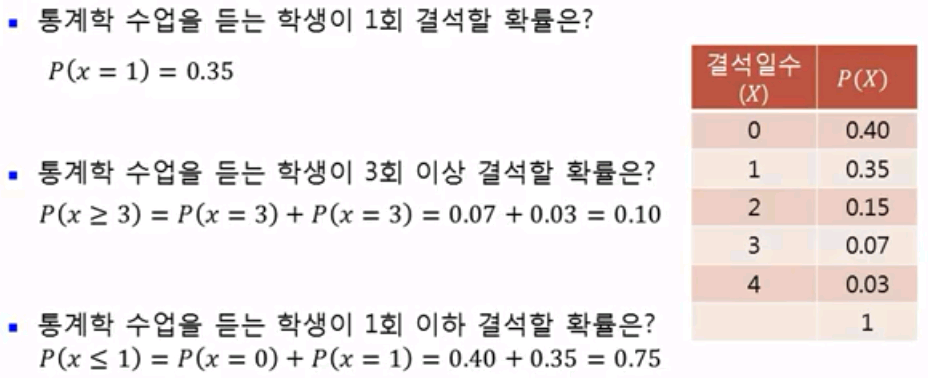


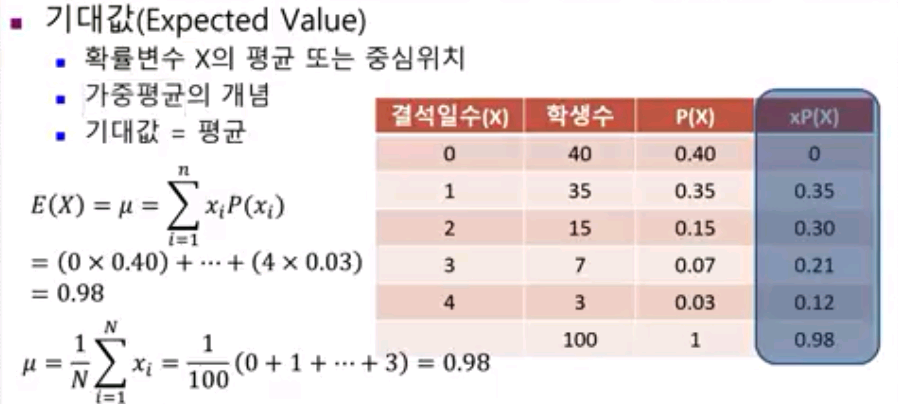
Ch06 이산 확률분포

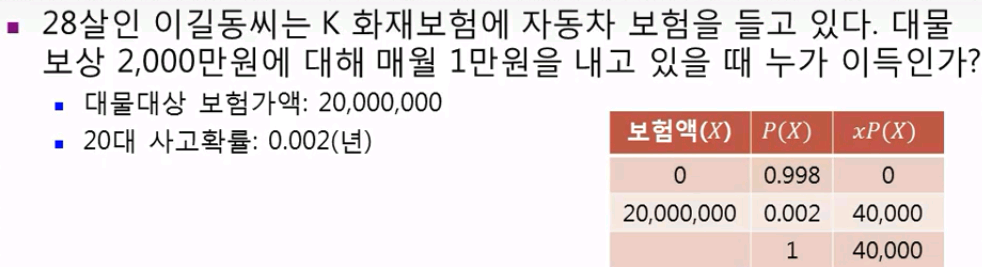
* 확률변수(random 변수)
  + 표본 공간에 있는 값을 숫자로 변경. 분포를 만들어 주기 위해서
  + 확률변수 = X, 확률 변수의 값 = x
    - 표본공간: S = {HH, HT, TH, TT} 🡺 확률변수 X = {2, 1, 1, 0}
  + **이산확률변수**(“별개의,”분리된”)
    - 특정한 수치만을 가지는 확률변수(정수)
    - 불량품 수, 방문자 수
  + **연속확률변수**
    - 어떤 범위에서 연속적인 값을 취하는 확률변수(실수).
    - 체온, 몸무게, 통근시간, 전구의 수명
* 확률분포(probability=가능성)
  + 확률 분포란 **확률변수의 모든 값과 그에 대응하는 확률들이 어떻게 분포**하고 있는지
  + 확률변수 X의 각 값()에 대응하는 확률을 표시
    - 이산확률분포
      * 일양분포, 이항분포, 포아송분포, 초기하분포, 기하분포
    - 연속확률본포 평균분포(정규본포, t-분포), 분산분포(분포, f분포)
* 이산 확률분포







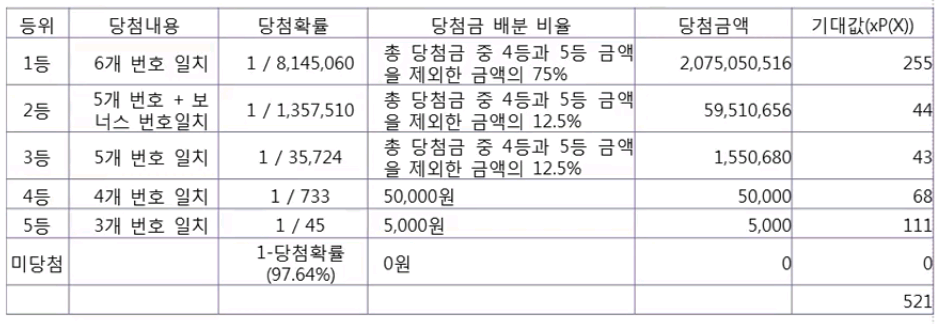




* 0 \*0.998 + 20,000,000 \* 0.002 = 40,000
* 회사수익 120,000 – 40,000 = 80,000
  + - * 보험 회사가 인건비, 사고처리비, 수익으로 30% 를 책정했다면 얼마로 보험료를 책정해야 하나?

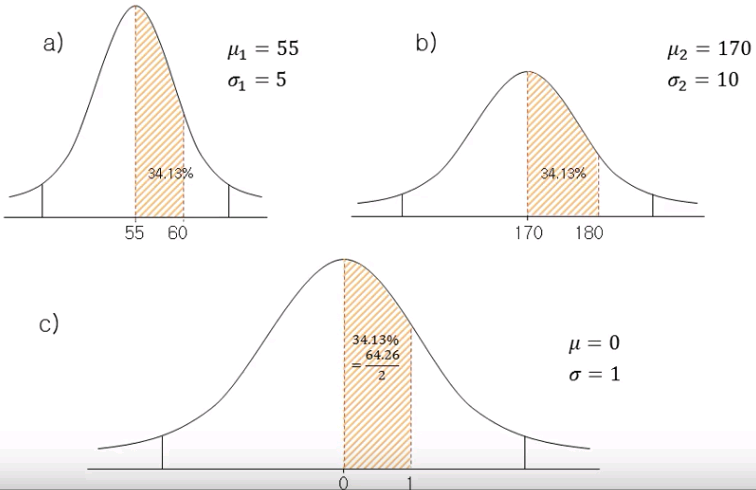
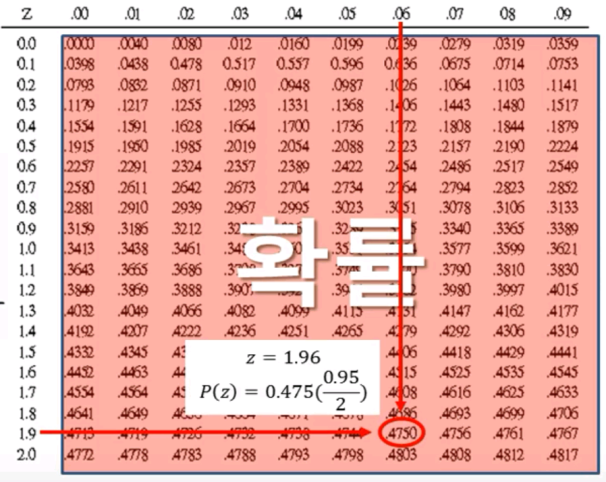
🡺 40,000 \*(40,000\*0.30) = 52,000

* + - * 로또번호 기대값 🡺 이익 = 521 – 1000 = -479

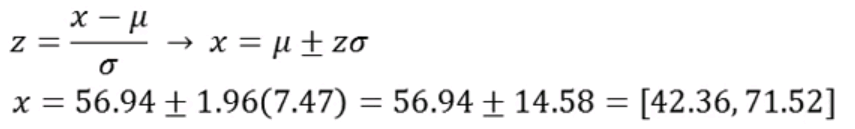


Ch07 연속확률분포

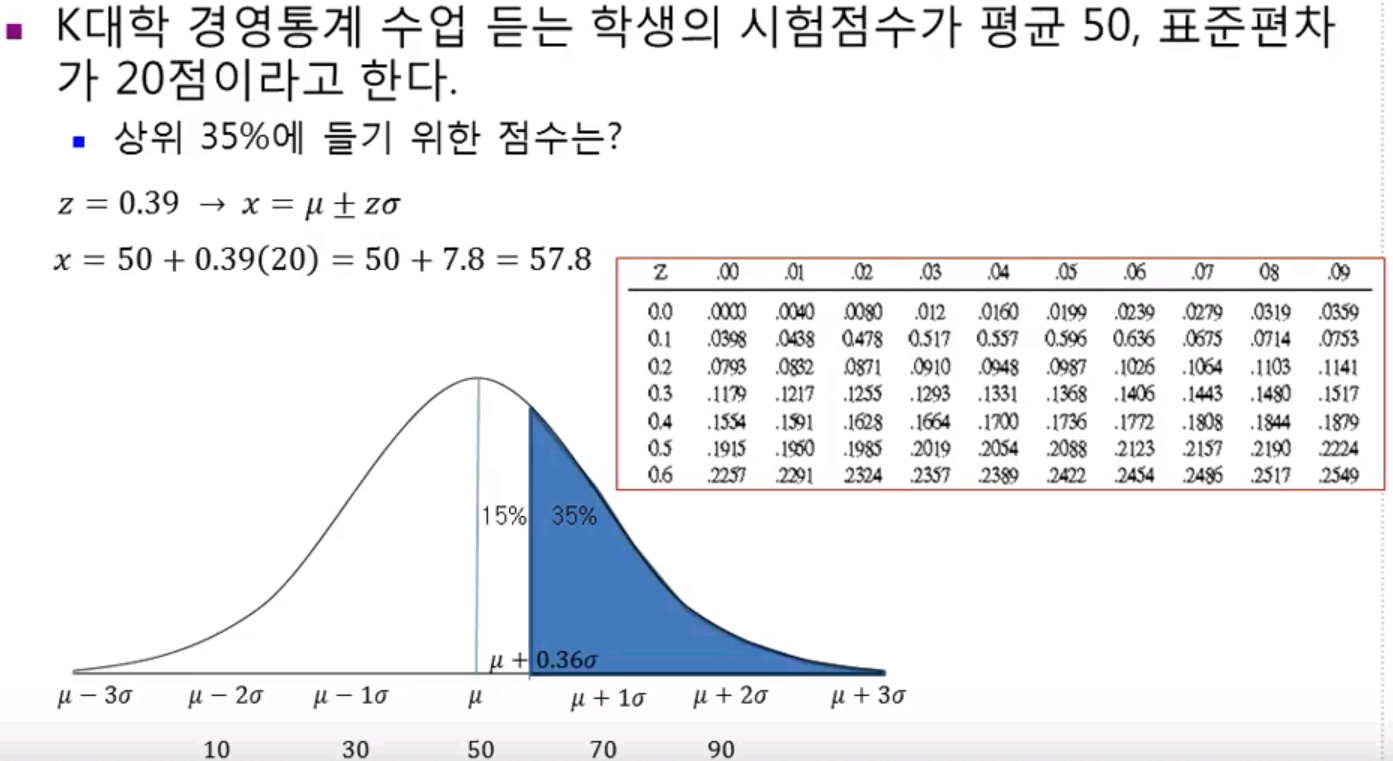
* 정규분포
  + 평균을 중심으로 좌우대칭이고 종모양을 갖는 확률분포
  + 샘플을 추출해서 모집단의 모수를 예측할 때 이용.
  + 통계분석 시 모집단의 분포를 정규본포라고 가정하고 통계분석을 한다. (극한중심)
  + 표준정규분포
    - **평균을** 0으로 표준화한 값으로부터 특정 **표준편차 값(Z)**에 대응하는 **확률**

* + - 표준확률변수
      * 표준편차의 Z값
      * 측정단위 등과 관계없이 자료를 표준화 시킨 값.
  + 정규분포표로 계산방법
    - **평균**과 **표준편차**와 **표준정규분포표**(평균과 표준편차에 따른 확률 표: Z값)를 이용

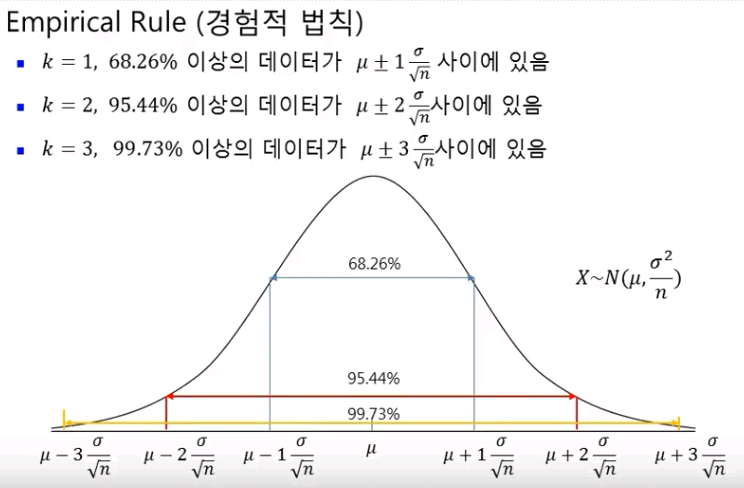


 %

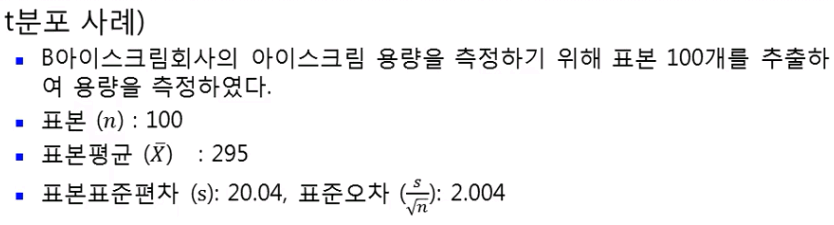
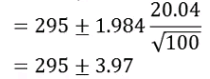


Ch08 통계적 추론

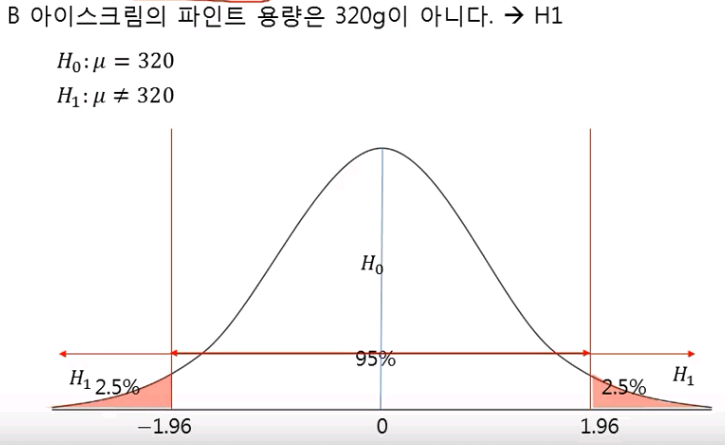
* 표집분포
  + 모집단에서 일정한 크기 (n)으로 표본을 뽑아서 각 표본의 평균을 계산하였을 때, 그 표본의 평균 들의 확률분포
  + 표본오차 sampling error – 표본 과 실제 모수와의 오차
  + 표준편차 standard deviation - 평균으로부터 퍼진 정도.
  + **표준오차 standard error –** 표본평균들의 표준편차. 표본 표준편차를 원소 개수의 제곱근으로 나누어준 값.
* 중심극한 정리



* 통계적 추론 – 추정, 가설검정
* Ch08-1 추정
* 추정
  + 모수 값이 속할 것으로 기대되는 일정한 범위(신뢰구간)을 이용하여 모수를 추정
  + 표본오차가 존재
  + 신뢰구간 – 신뢰수준의 확률 95%. 동일한 모집단에 대해서 동일한 방법으로 표본을 다시 뽑아서 신뢰구간을 구하게 되면 100번 중 95번은 모수를 포함한다.
  + 방법
    - 모집단의 표준편차를 알 경우: 표준정규분포
    - 모집단의 표준편차를 모를 경우: student t 분포. 표본의 크기가 30보다 크면 분포는 정규분포에 접근

* Ch08-2 가설검정
* 가설검정
  + 모집단 모수의 값을 설정하고 (가설설정), 표본 통계치를 통해 확률적으로 진위를 검증하는 방법
  + 유의수준: 모수와 통계량의 차이가 커서 확률적으로 가설을 기각할 수 있는 값. 5%
  + 종류: 귀무가설, 연구가설
    - H0 귀무가설 – 기존에 알려져 있는 사실 **(통계적 검정 대상)**
    - H1 대립가설 – 새로운 사실, 현재 믿음의 변화가 있는 사실, 뚜렷한 증거로 입증하려고 하는 주장
  + **통계적 가설검정 - 귀무가설을 받아 들일 것 인지, 아니면 기각할 것인지를 표본을 구해서 검증**
  + H0은 현재 알려진 사실이기 때문에 명확히 가설로 진술할 수 있기 때문에 귀무가설만 검증
  + H0 가설이 맞는 확률(95%)로 검증해서 귀무가설이 받아 들일 것인지 기각할 것인지 결정한다.

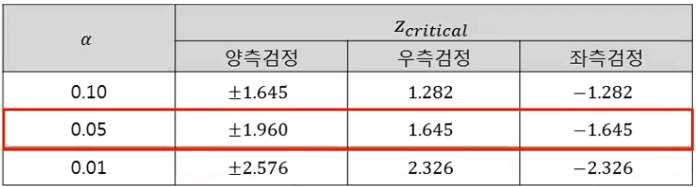


* + 가설검정의 오류
    - 100% 완벽한 증거를 수집할 수 없어서 오류(error)가 존재
    - 잘못된 의사결정을 내릴 수 있는 위험수준(risk)를 결정해야 함.

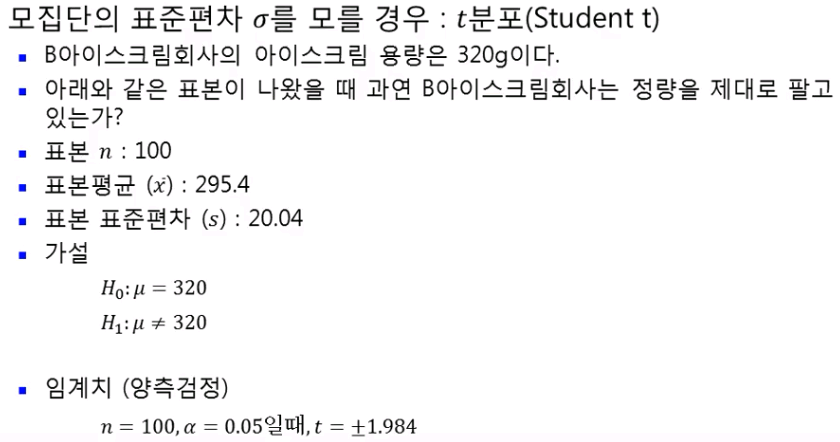


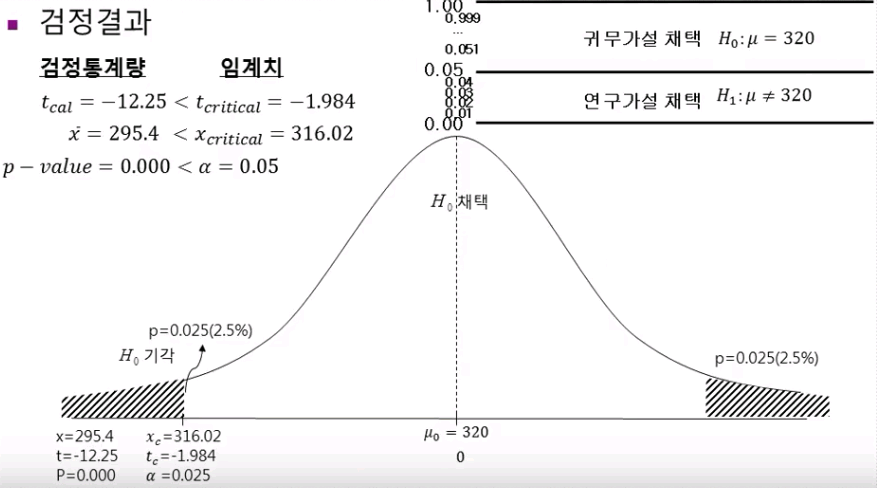
* + - * 1종 오류: 맞았는데 틀렸다고 판명할 오류
      * 2종 오류: 틀렸는데 맞았다고 할 오류
  + 유의수준

H0이 진실인데 H0를 기각할 수 있는 오류를 범할 확률의 최대 허용치 0.05

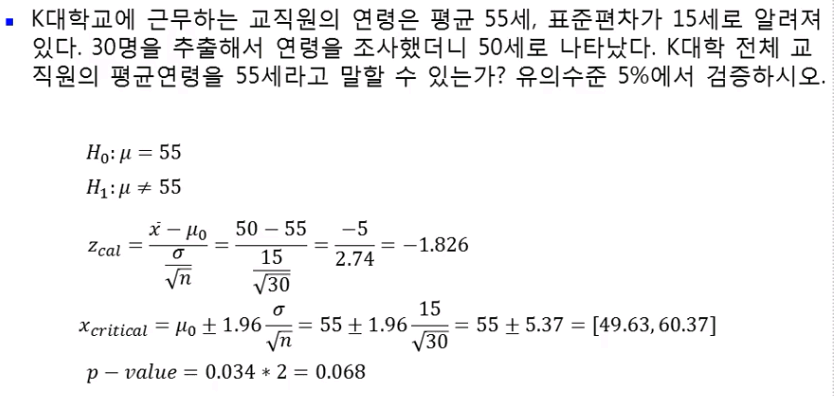


* + 유의확율(p-value)
    - 귀무가설이 맞는다는 가정하에서 표본 통계량의 값이 나타날 확률
    - p-value < 0.05 🡺 귀무가설 기각 . 통계 패키지에서 p-value을 계산해 줌
* 모평균 가설검정





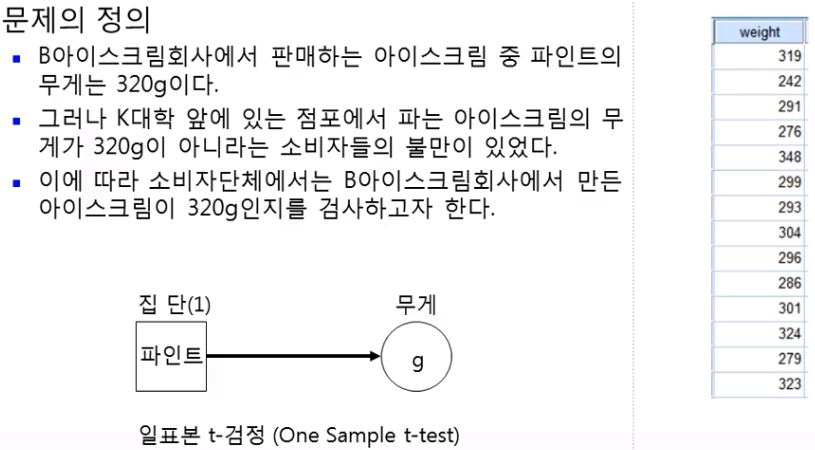
\*\* 연습 \*\*



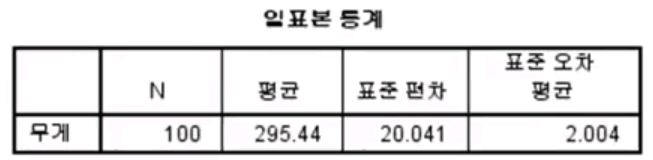
Ch09 통계분석방법

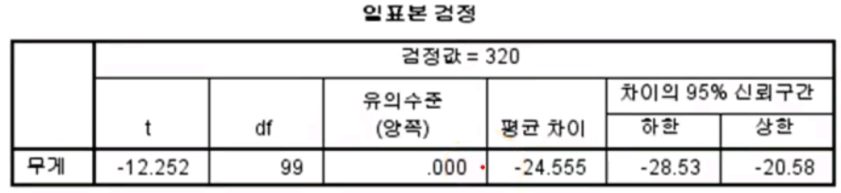


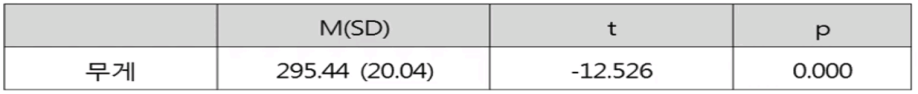
* + 차이검정 – 그룹 수
  + 관계검정 – 그룹X, 변수와 변수간의 관계(영향을 주는 요소)
    - Independent sample t-test 🡺 One Way ANOVA
    - Paired Sample t-test 🡺 Repeated Measured ANOVA



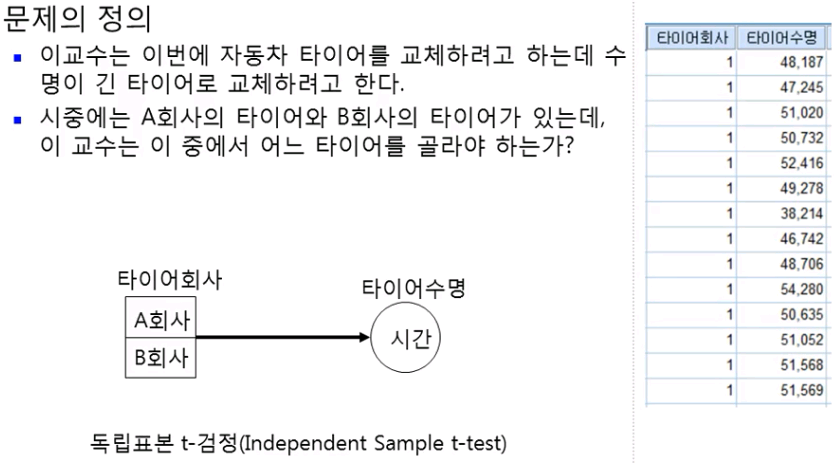
* + 가설 H0 = 320, H1 != 320



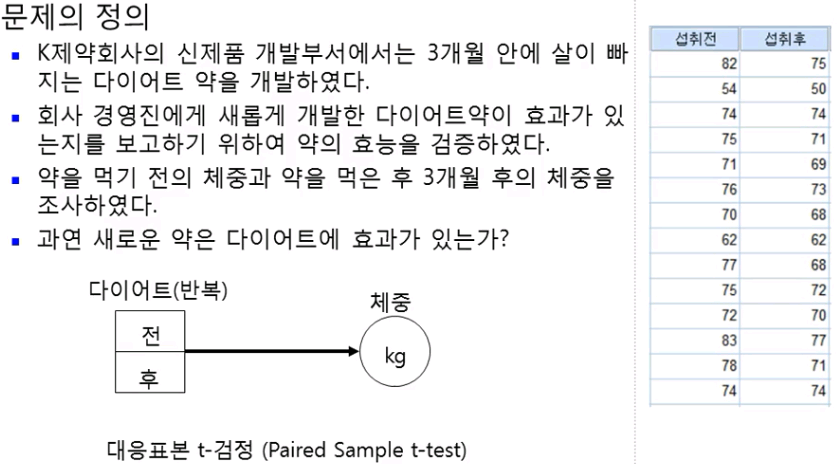


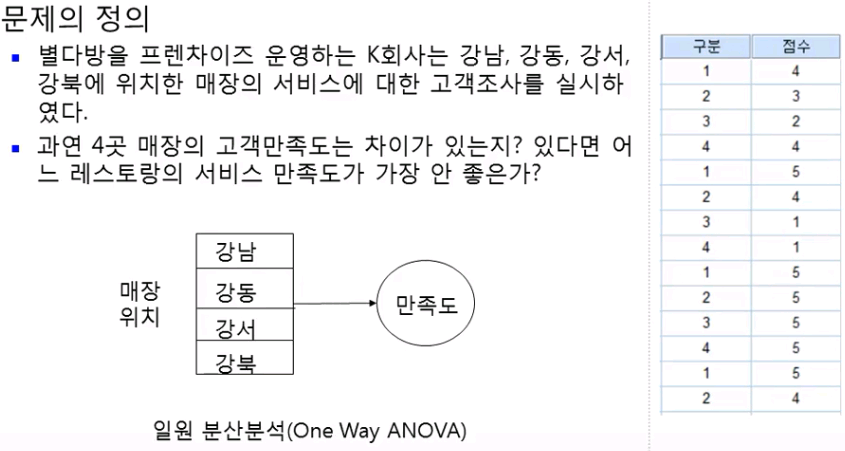


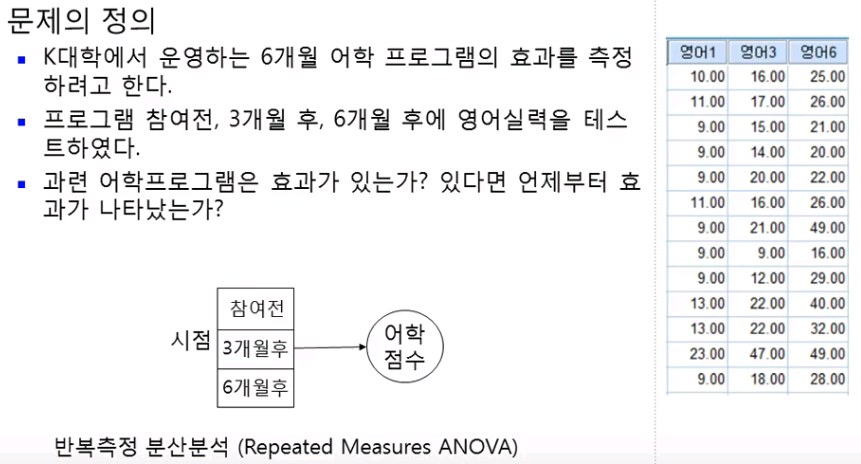
* Independent Sample t-test



* Paired Sample t-test (대응표본)

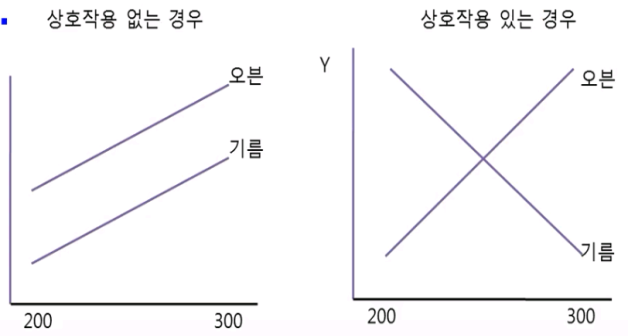


* One Way ANOVA  
  
* Repeated Measure ANOVA

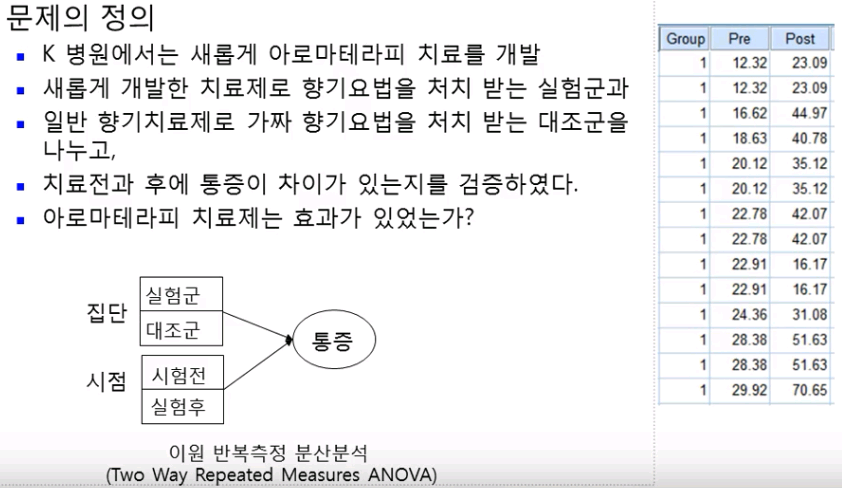


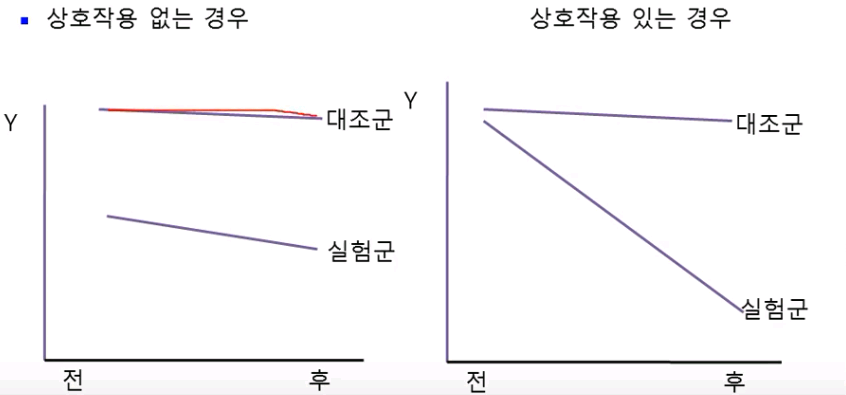
* Two Way ANOVA



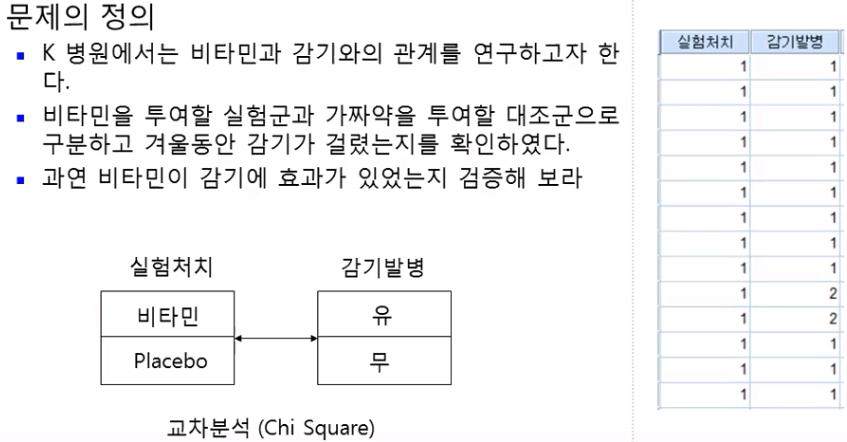


* Two Way Repeated Measures ANOVA

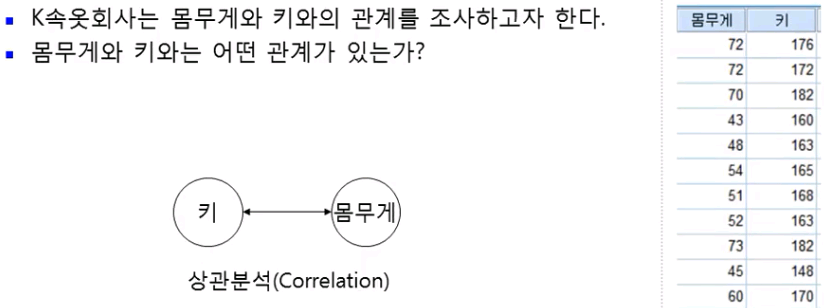




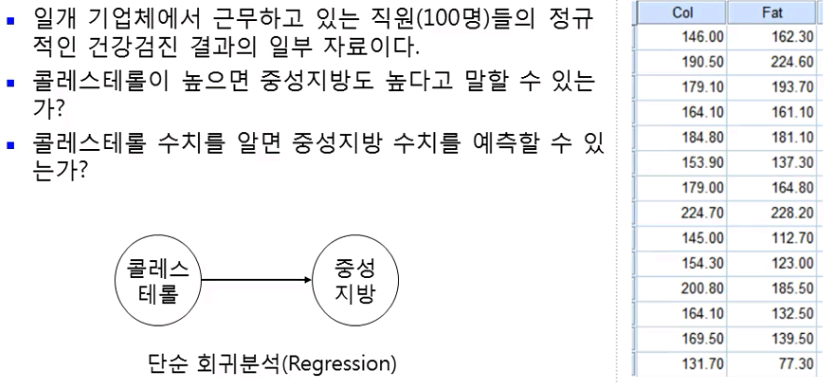
* 교차 분석 Chi-Square test
  + 범주형 변수간의 관계검정



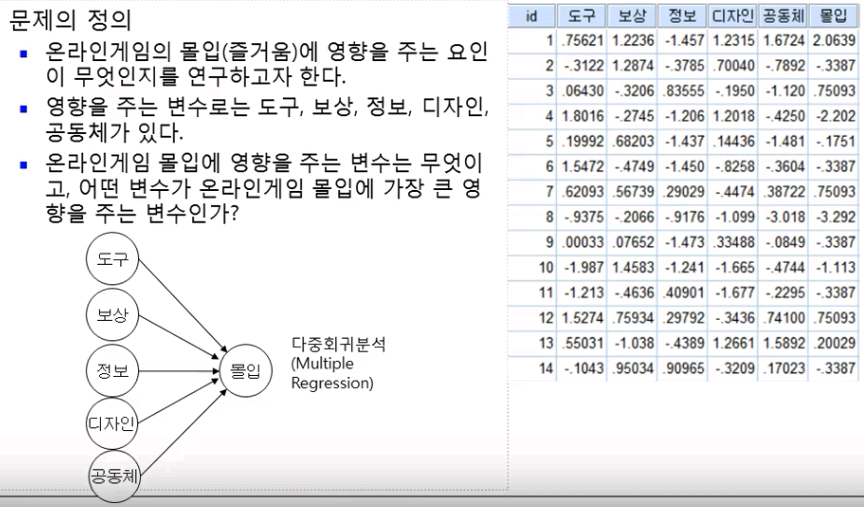
* 관계검정 (Correlation Test )
  + 두 변수의 형태가 둘 다 범주형이거나 둘 다 수치형
  + 관계 1:1
  + 서로가 관계는 있는데 영향(원인과 결과)은 없다.



* 단순회기분석 (Regression) 거의 없음
  + 관계 1:1
  + 원인과 결과가 있다.



* 다중 회기분석 (Multiple Regression)
  + 관계 n:1
  + 원인과 결과가 있다



* Logistic Regression
  + 자료의 형태가 달라
  + 결과 변수가 그룹으로 분류

