목차

[Ch01 통계 1](#_Toc5228115)

[Ch02 표본조사 2](#_Toc5228116)

[Ch03 자료 2](#_Toc5228117)

[Ch04 기술통계분석 5](#_Toc5228118)

[Ch05 확률과 카운팅(counting) 9](#_Toc5228119)

[Ch05\_A 확률규칙 11](#_Toc5228120)

[Ch06 이산확률분포 12](#_Toc5228121)

[Ch07 연속확률분포 12](#_Toc5228122)

[Ch08 포본분포와 추정 12](#_Toc5228123)

[Ch09 가설검증 12](#_Toc5228124)

[Ch10 통계분석방법 12](#_Toc5228125)

[Ch11 평균차이검정 12](#_Toc5228126)

[Ch12 분산분석 12](#_Toc5228127)

[Ch13 회귀분석 13](#_Toc5228128)

[Ch14 교차분석 13](#_Toc5228129)

Ch01 통계

* 통계란 --**자료**를 분석해서 자료를 **정보화** 시키는 것
  + 관심대상에 대해 관련 **자료를 수집**하고

🡺 **표본추출**(sampling) 🡪 확률

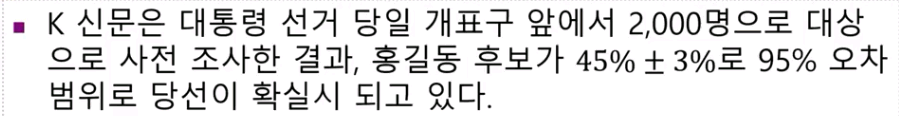
* + **자료를 요약 정리**

🡺 **기술통계학**(descriptive statistics)

* + 불확실한 사실에 대한 결론이나 일반적인 규칙성을 추리

🡺 **추리통계학**(inferential statistics)

* 기술통계학
  + 수집된 자료를 정리, 요약하여 수치, 표, 그래프로 자료의 특징을 파악
  + 자료가 가지고 있는 원래의 특성만 표현
  + 평균 키는 얼마인가? 대통령선거에서 각 후보가 받은 득표수.
* 추측통계학
  + 모집단의 일부인 표본을 분석하여 모집단에 대해 확률적으로 판단(추측)하고 일반화
  + 새로운 가설이 맞는지 틀리는지를 검증
  + 5000개 형광등에서 50개 샘플로 조사한 형광등 수명으로 불량품 예측





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

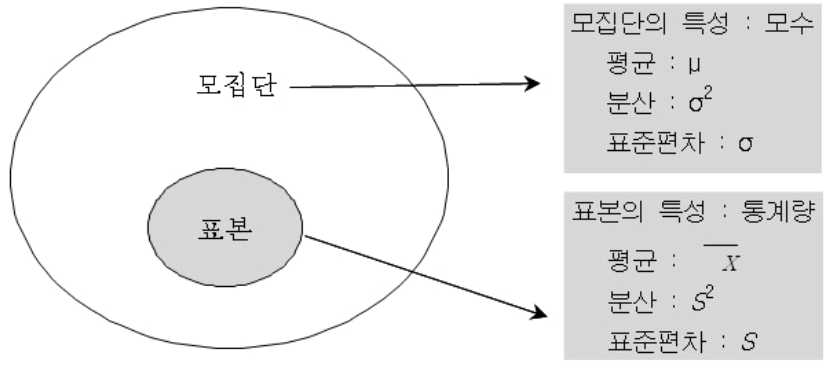
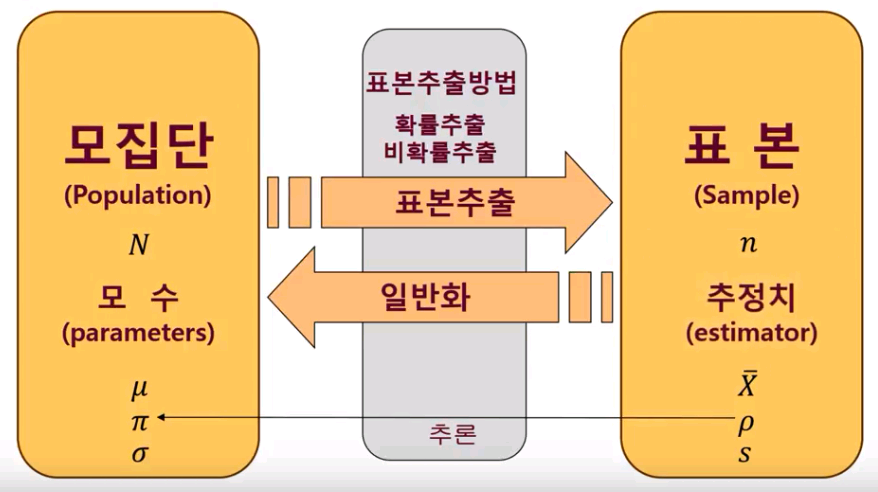
전수조사는 거의 불가능하고 매우 비효율적.

그래서 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)

* 표본오차 (sampling error)
  + 모집단에서 표본을 추출해서 조사하기 때문에 모수와 표본 통계량 사이에 생기는 오차
  + 표본의 크기를 클수록 표본오차감소
  + 표본오차의 허용범위를 확률로 구하는 것이 바로 가설검정
* *실제 알고자 하는 값은 모집단의 모수(특성)인데 전수조사가 불가해서 표본조사를 하게 된다. But 실제 모수와 표본조사의 통계량과는 오차가 발생함으로 그 오차범위를 확률을 이용해서 구하여 전체 값을 예측 하게 된다.*

Ch03 자료

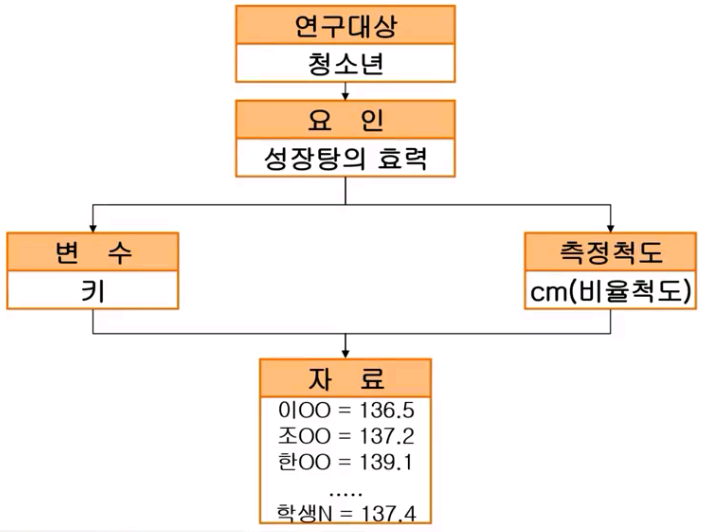
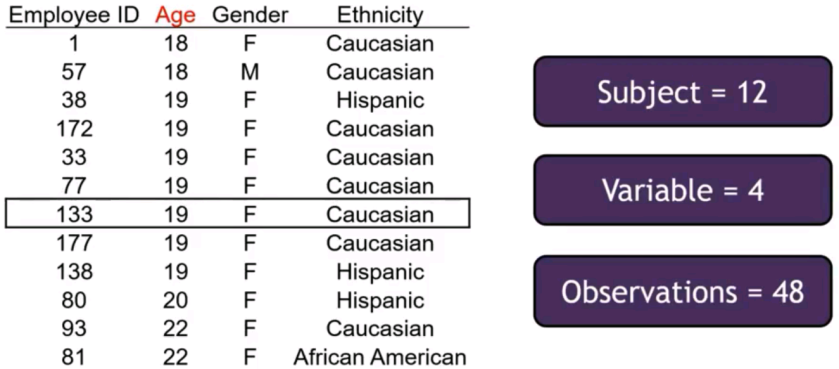
* 자료



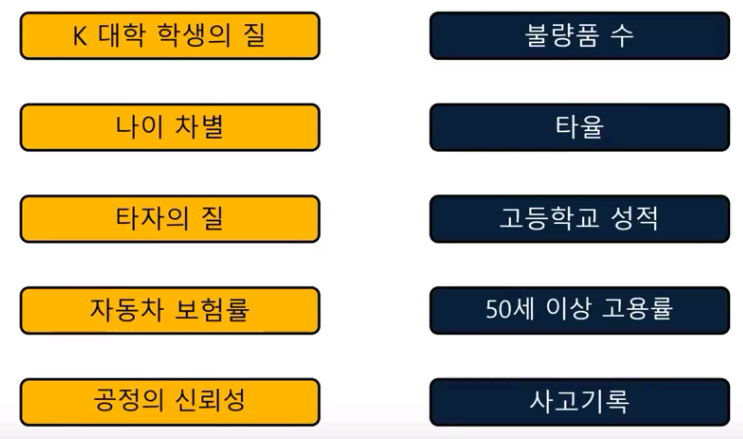
* + 개체(subject) – 연구자가 관심을 갖는 **대상**.
  + 요인(Factor) 또는 개념(concept) – 개체에 대한 특성 중 연구자가 관심을 갖는 **특성**.
  + 변수(Variable) – 요인을 구성하고 있는 요소.
    - 변수는 일정한 측정 단위로 표현될 수 있어야 한다. 특정 단위 또는 숫자로 표현이 가능해야 한다.

신장의 경우 cm, 몸무게는 kg으로 혈액형은 A, B, O

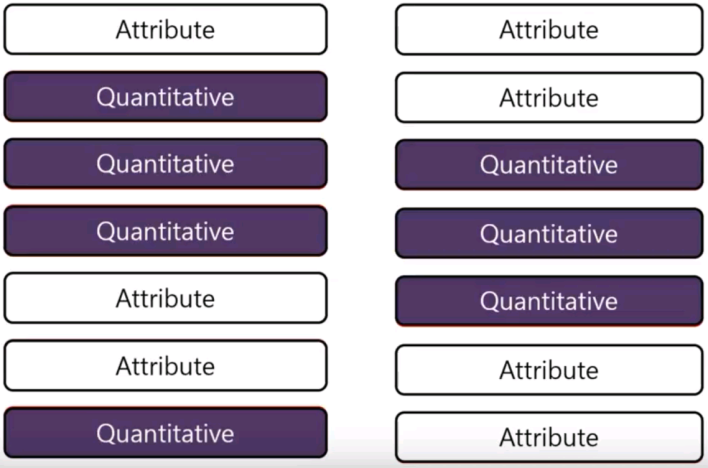
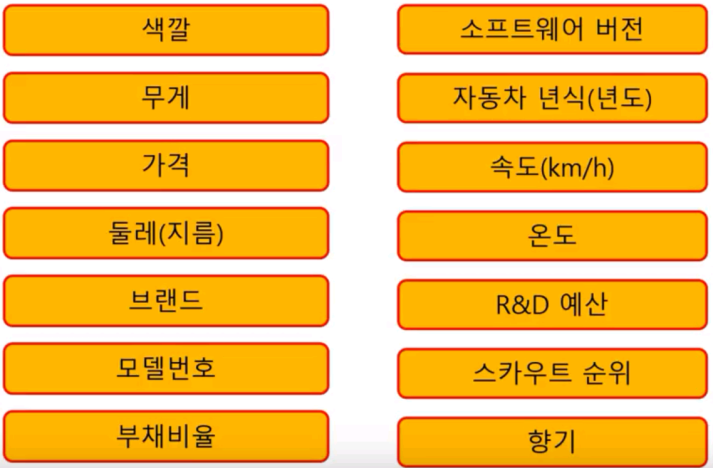
* + 척도(Measurement)
  + 자료(Data) – Observations

* 요인(Factor) VS 변수(Variable)

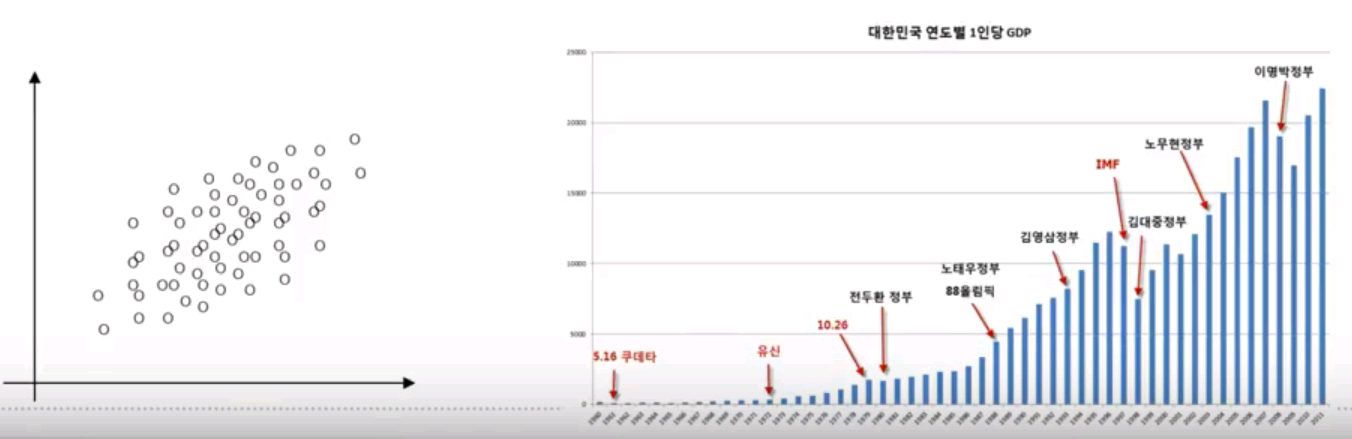


* 자료의 종류 10가지 분석기법은 자료에 따라 분석방법이 결정된다. 자료의 종류를 구별할 수 있어야 한다.
  + 질적 자료 (qualitative data, categorical data)
    - 명목(categorical), 속성 (attribute) 자료 또는 변수
    - 몇 개의 특성에 의한 범부를 나누어 코드 숫자로 나타낸 자료
    - 성별(남자=1, 여자=2) 연산할 수 없음.
    - 남자,여자: 평균1.5(X), 퍼센트/비율(O)
    - \*\* 학년, 경영학과, 혈액형, 연령대 – 카테고리
  + 양적 자료(quantitative data, Numerical data)
    - 연속 자료 변수 (continuous variables)
    - 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
    - \*\* 양적변수는 질적변수로 변경가능.

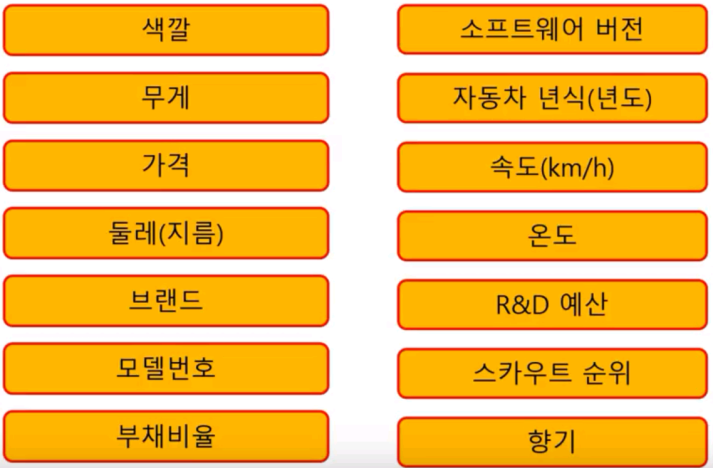
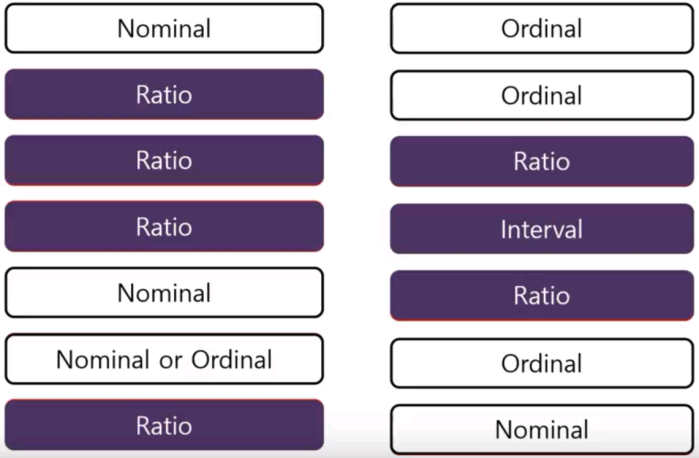


* 자료의 종류(시간 관점)
  + 종단 자료(Cross-sectional data) – 한번의 시간에 얻어진 데이터
  + 횡단 자료(Longitudinal data) – 여러 시간에 걸쳐 얻어진 데이터 Ex. 주식, 2010 ~ CDP변화

Ex. 3월22일 삼성전자 주식가격? 양적. 종단



* 척도의 종류
  + 척도(Measurement )의 종류
    - 명목척도(**Nominal**)
      * 명목 척도는 숫자로 표현될 수 있지만 수량적인 의미를 갖지 않고, 범주(카테고리)를 구분하는 용도로 쓰이는 척도
      * 성별(남/여) 국적(한국, 일본, 중국 등), 직업(회사원, 공무원, 자영업 등)
    - 서열척도(**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

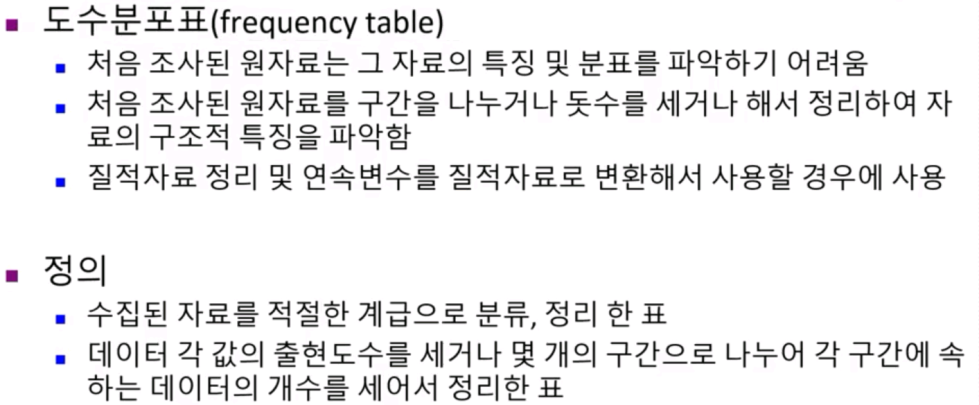
기술 통계에서 "기술"은 영어로 Descriptive 입니다. 한국말로는 "묘사하는" 또는 "그려서 설명하는"이라는 뜻입니다. 즉, 기술 통계는 우리가 수집한 데이터를 묘사하고 설명하는 통계 기법들을 말합니다.

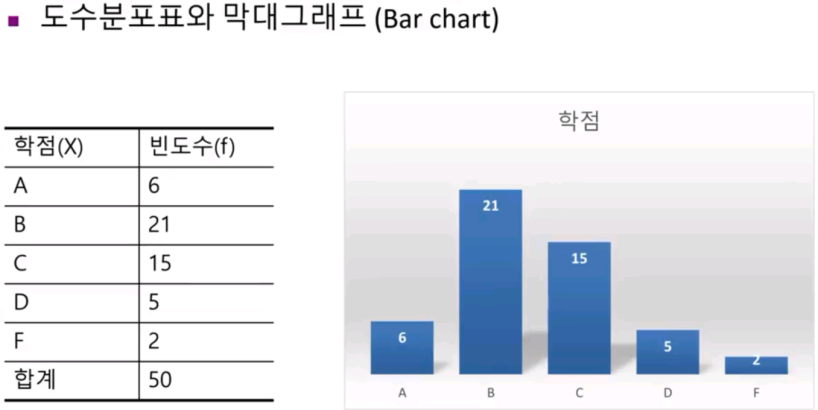
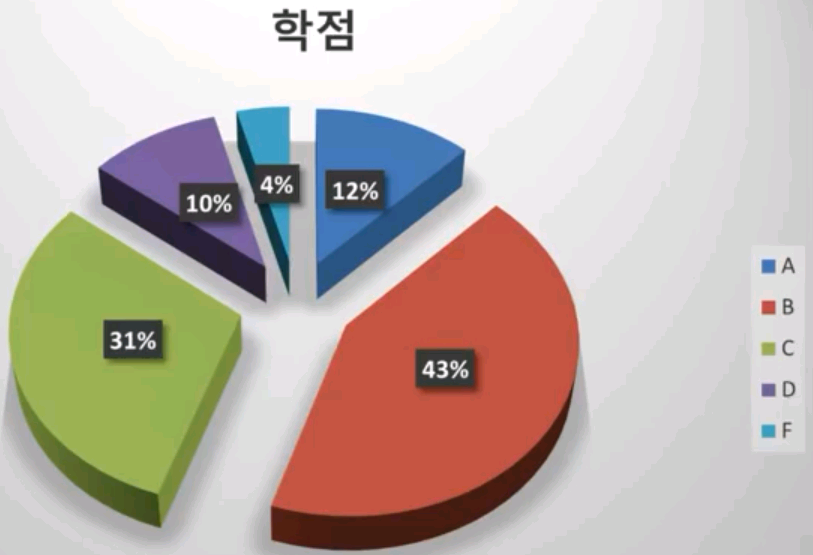
그렇다면 기술 통계에는 어떤 것들이 있을까요? 다시 말해 **수집한 데이터를 설명하는 통계 기법**이 무엇이 있을까요? 기술 통계 기법은 크게 또 두 가지로 구분할 수 있습니다. 하나는 우리가 수집한 데이터를 대표하는 값이 무엇인지 또는 어떤 값에 집중되어 있는지를 다루는 기법입니다. 조금 전문적으로 말씀드리면 데이터의 **집중화 경향 (Central tendency)에 대한 기법**이라고 말할 수 있습니다. **평균 (mean), 중앙값(median), 최빈값(mode)** 등이 바로 집중화 경향에 속하는 것들 입니다.

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 있는를 설명하는 방법입니다. 대표적으로 **표준편차 (standard deviation), 사분위(quartile)** 값 등이 있습니다.

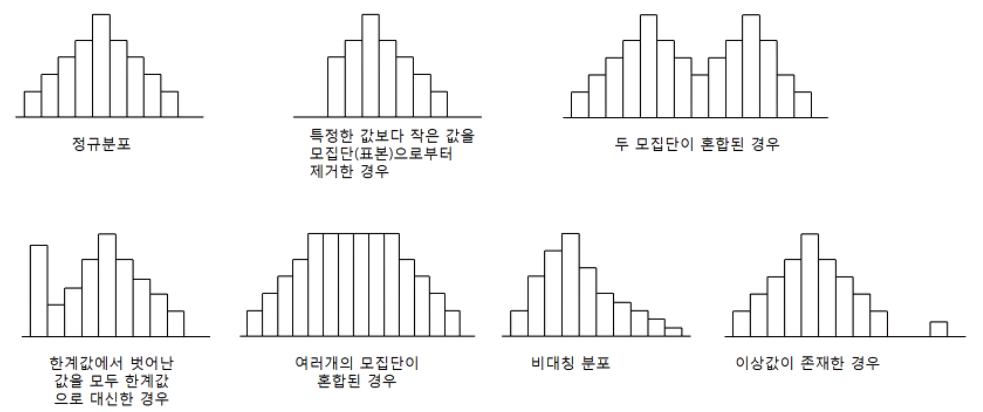
우리는 기술 통계 기법을 통해서 수집한 데이터의 전체적인 모양을 그릴 수 있습니다. 예컨대, 우리나라의 국민 1인당 평균 소득이 2만 달러라고 합시다. 이 값은 우리나라 국민의 소득 수준의 대표값입니다. 하지만 대표값만이 중요할까요? 아닙니다. 분산도 중요합니다. 예를 들어 국민 1인당 평균 소득이 아무리 높아도, 소득에 대한 편차도 함께 높은 값을 가지고 있다고 가정해 봅시다. 편차가 크다는 것은 소득 분포가 넓게 분표되어 있다는 뜻이고, 이는 다시 말해 국민의 소득 편차가 크다는 뜻입니다. 이를 통해 "소득의 분배가 잘 이루어지지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

* 도수분포표 & 히스토그램

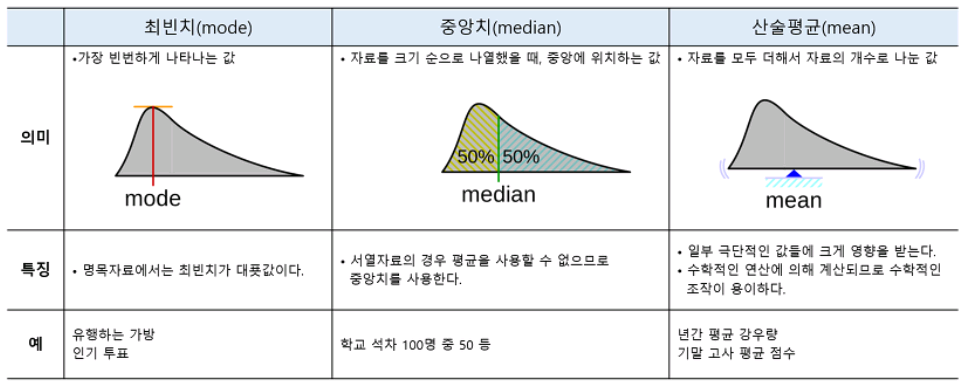
 

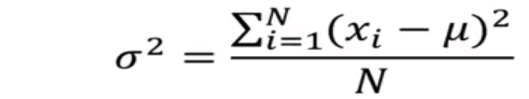
* 히스토그램
  + 도수 분포표를 시각적으로 표현한 막대 그래프
  + 히스토그램의 x축(가로축)은 구간을 나타내고, y축(세로축)은 각 구간별 빈도수를 나타낸다
  + 빈도수로 히스토그램의 모양을 결정됨으로 데이터가 어떻게 생겼는지 한 눈에 볼 수 있다.



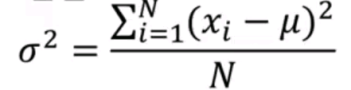
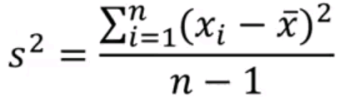
* 중심화 경향
  + 수집한 자료 전체를 대표하는 값이 무엇인지 나타내는 통계.
  + 최빈값: 수집한 데이터 중 그 빈도가 가장 많이 나타나는 데이터 (명목 자료의 대표값)
  + 중앙값: 자료를 크기 순으로 정렬했을 때, 중앙에 위치하는 값 (순위 자료의 대표값)
  + 평균값: 자료를 모두 더해서 전체 자료의 갯수로 나눈 값 (정량적 자료의 대표값)



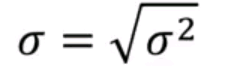
* 평균의 문제
  + 평균값은 이상 치에 의해 민감하게 영향을 받는다. 실제로 자료가 어떻게 변화하는지 알 수 없다.
* 분산(variance)
  + 자료가 평균을 중심으로 얼마나 광범위하게 분포하고 있는 가를 하나의 수치로 나타낸 통계량
  + 평균은 모든 데이터의 중심이니까 거리를 다 더하면 0이 되기 때문에 +, - 를 없애려고 제곱해서 루트



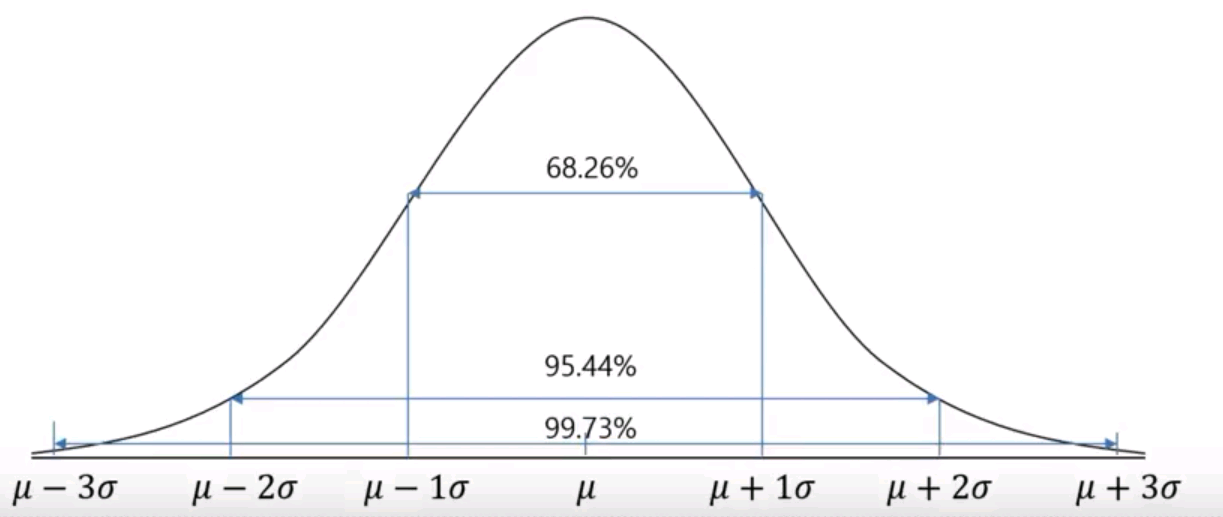
* 표준편차(standard Deviation) – 평균으로부터 자료가 흩어진 정도
  + 분산

* + 표준편차

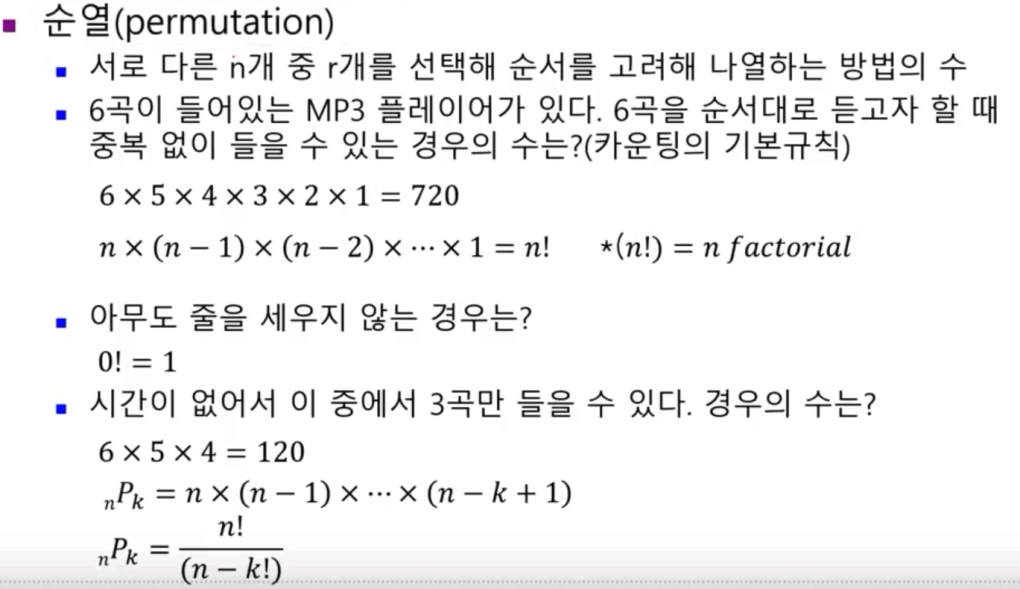
 

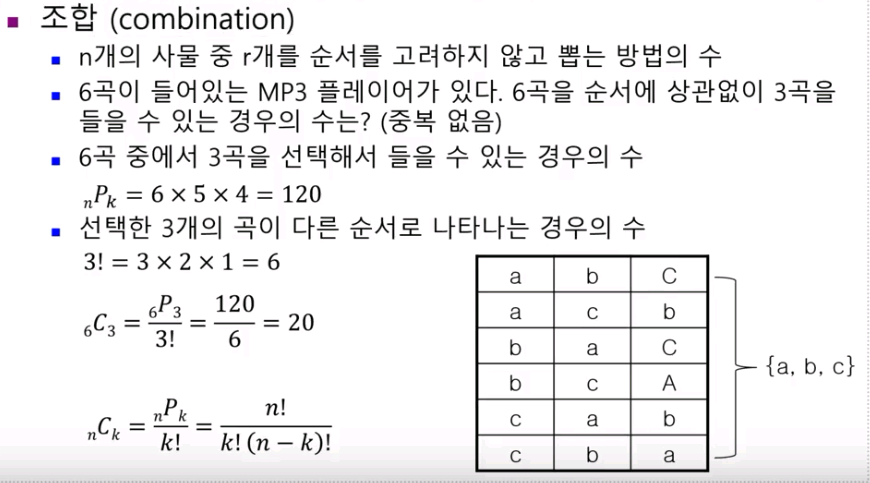
* 표준화된 자료
  + Empirical Rule (경험적 법칙)

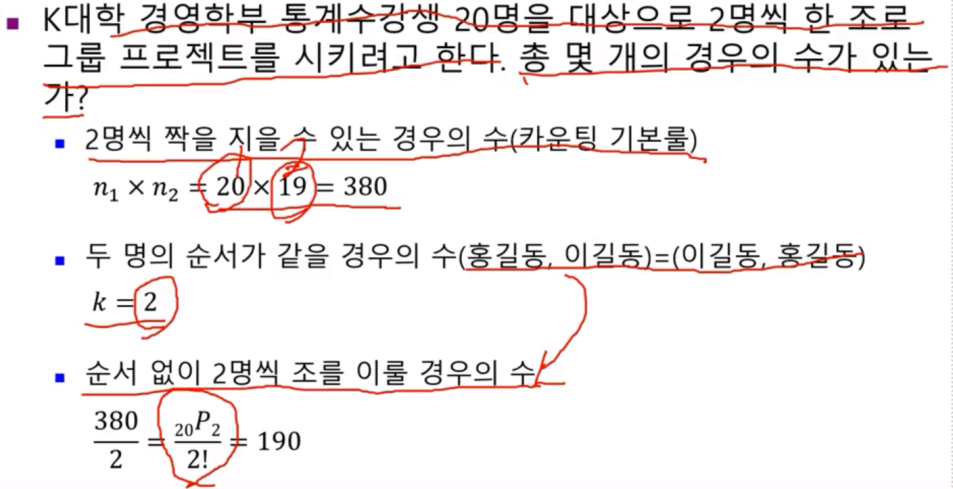


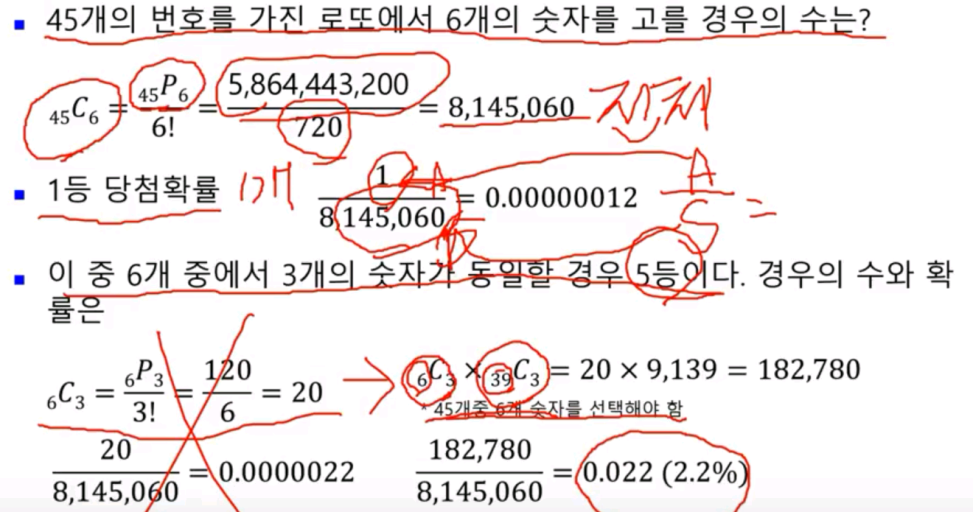
* + - 68.26% 이상의 데이터가 사이에 있음
    - 95.44% 이상의 데이터가 사이에 있음
    - 99.73% 이상의 데이터가 사이에 있음

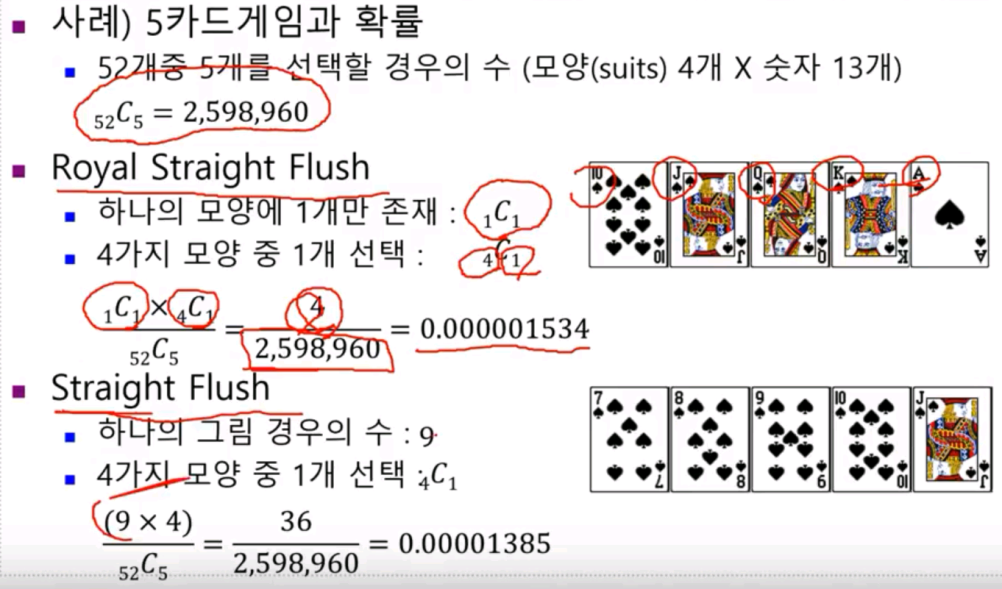
Ch05 확률과 카운팅(counting)





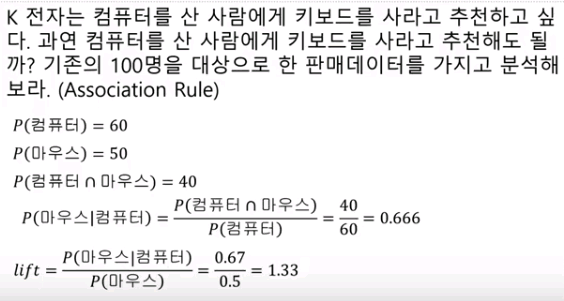






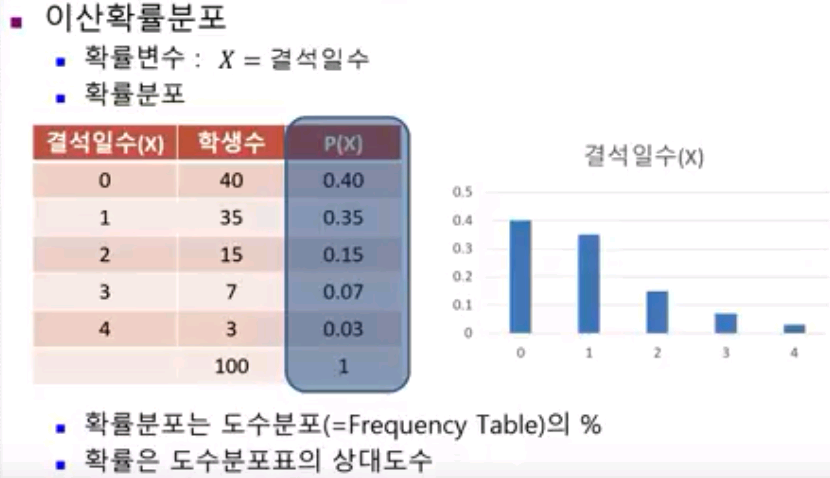
Ch05\_A 확률규칙

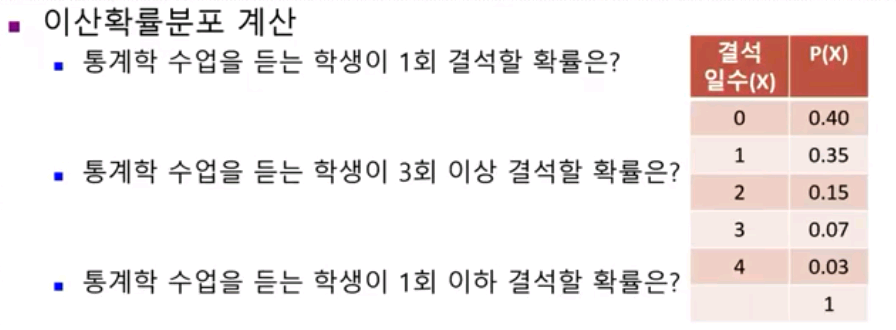
* 확률규칙
  + 확률(probability) & 승산비(odd ratio)
    - Probability - 전체 중에서 사건이 차지하는 크기. 1을 넘을 수가 없음.
    - Odd ratio - 사건 A와 B의 상대적 크기. 성비, 인구밀도. 1을 넘을 수가 있지
    - 총인구52,341,015, 여자 25,651,015, 남자 25,663,668
      * 여자비율=25,651,015/52,341,015=49.99
      * 여성비= 25,651,015/25,663,668=1.00
  + 조건부확율(conditional probability)
    - A라는 조건이 주어진 상태에서 B가 발생할 확률. 교차분석, 데이터 마이닝의 연관성분석 중요한 개념
    - P(B|A) = P(A∩B)/P(A)

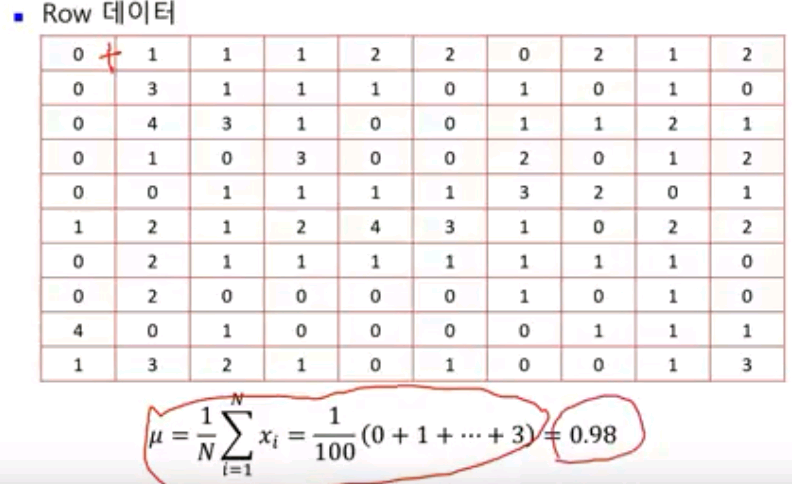


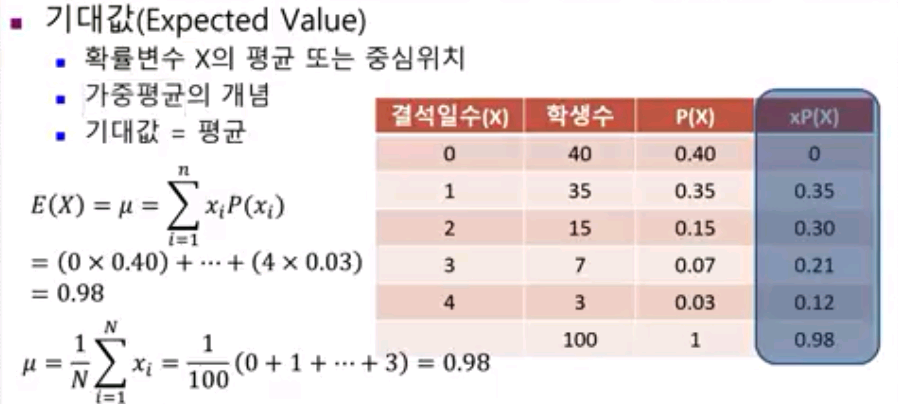
Ch06 이산확률분포

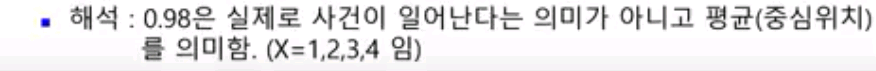
* 확률변수
  + 이산확률변수(“별개의,”분리된”)
    - 확률 변수 X가 어느 구간의 모든 실수 값을 택하지 않고 0,1,2...와 같은 고립된 값만을 택하는 확률변수(정수). 불량품 수, 방문자 수
  + 연속확률변수
    - 확률변수가 취하는 값이 어떤 범위에서 연속적인 값을 취할 수 있는 확률변수(실수).
    - 체온, 몸무게, 통근시간
* 확률분포
  + 확률 분포란 **확률변수의 모든 값과 그에 대응하는 확률들이 어떻게 분포**하고 있는지
  + 확률변수 X의 각 값()에 대응하는 확률을 표시
  + 이산확률분포
    - 일양분포, 이항분포, 포아송분포, 기하분포
  + 연속확률본포
    - 평균분포(정규본포, t-분포), 분산분포(분포, f분포)
* 확률분포표
  + 확률변수 x에 대한 확률을 도수분포표로 만든 것







****



Ch07 연속확률분포

* 정규분포
  + 샘플을 추출해서 모집단의 모수를 예측할 때 이용.
  + 통계분석 시 모집단의 분포를 정규본포라고 가정하고 통계분석을 한다.
  + 정규분포의 계산방법
    - 평균과 표준편차와 표준정규분포표(평균과 표준편차에 따른 확률 표: Z값)를 이용
  + 표준정규분포표
    - **평균을** 0으로 표준화한 값으로부터 특정 **표준편차 값(Z)**에 대응하는 **확률**
    - 표준확률변수
      * 표준편차(걸음걸이)에 Z값
      * 측정단위 등과 관계없이 자료를 표준화 시킨 값.
      * empirical rule z = x - μ / 🡺 x = μ
* 통계학의 목적 = 오차범위를 구한다.
  + 알고자 하는 값은 모수이나 전수조사 x
  + 표본의 평균을 구한다. 표본의 평균과 실제 값의 차이를 확률을 통해서 예측한다.
  + 95%로 믿을 수 있는 오차범위의 값을 구한다. 100명중 95명이 확신할 수 있는 값. 양쪽 끝의 오차범위를 구한다.

Ch08 포본분포와 추정

Ch09 가설검증

Ch10 통계분석방법

Ch11 평균차이검정

Ch12 분산분석

Ch13 회귀분석

Ch14 교차분석