문화 A0007

데이터사이언스입문

김 태 완

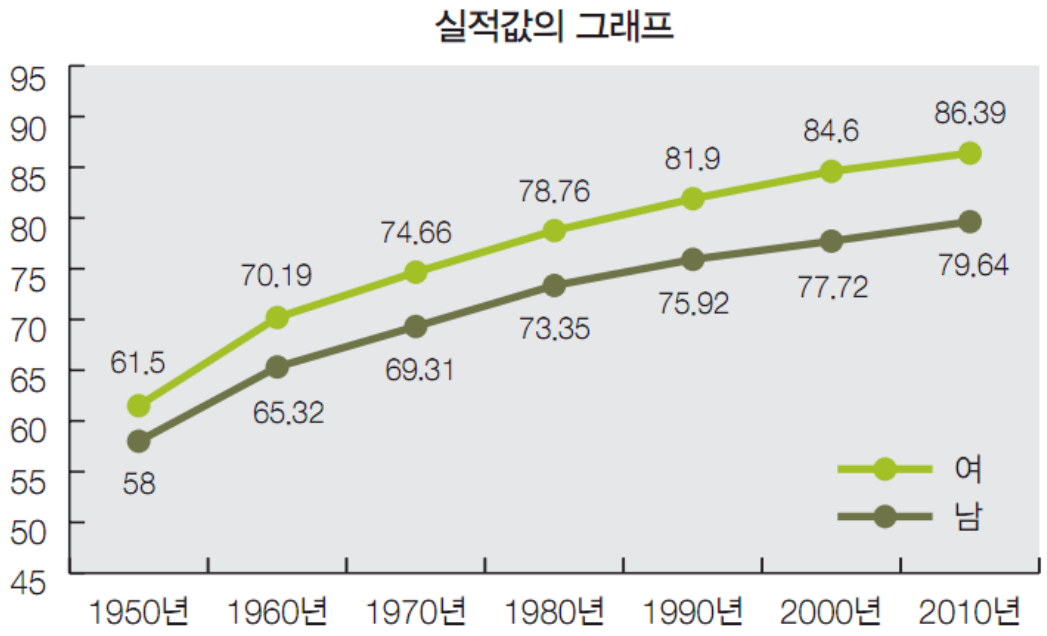
kimtwan21@dongduk.ac.kr

통계를 들어가기 전에...

- 데이터 정리

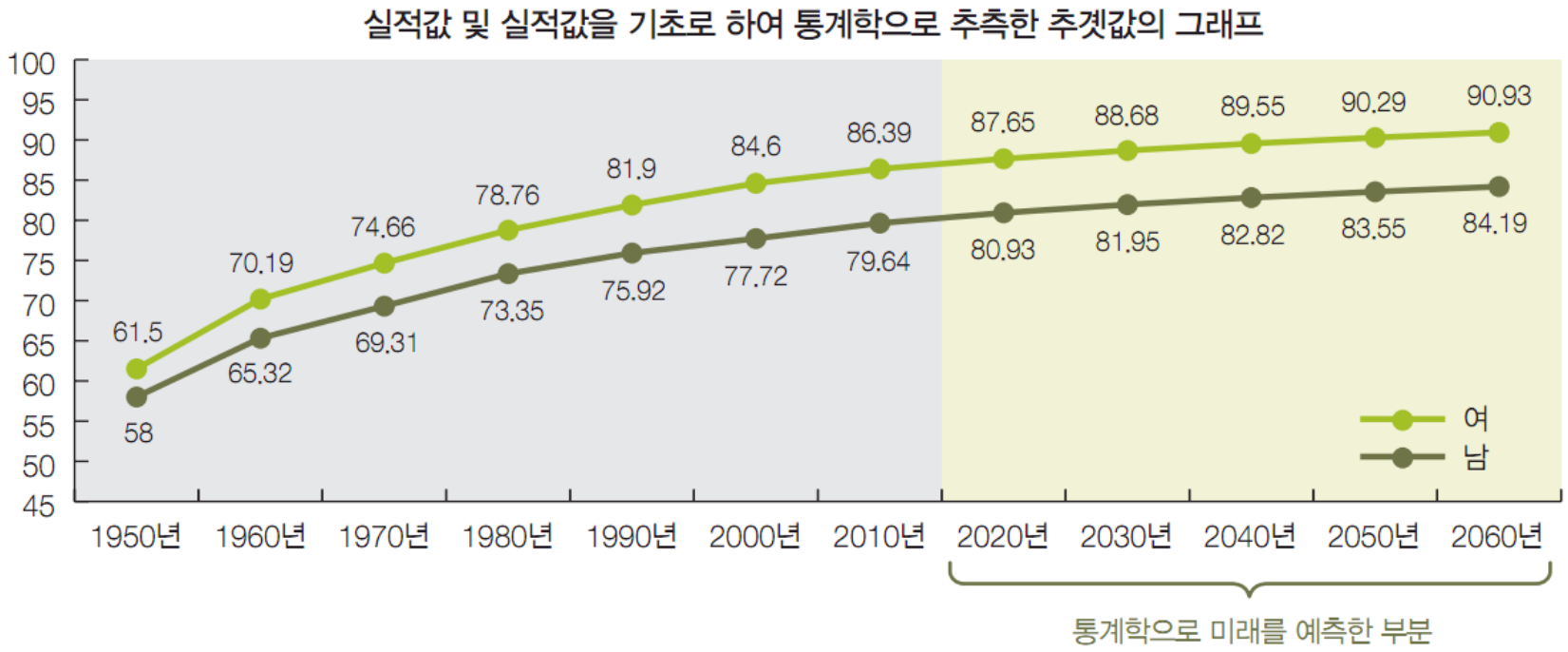
평균수명의 추이

		남	여
실적값	1950년	58	61.5
	1960년	65.32	70.19
	1970년	69.31	74.66
	1980년	73.35	78.76
	1990년	75.92	81.9
	2000년	77.72	84.6
	2010년	79.64	86.39
추계값	2020년	80.93	87.65
	2030년	81.95	88.68
	2040년	82.82	89.55
	2050년	83.55	90.29
	2060년	84.19	90.93



통계를 들어가기 전에...

- 데이터를 분석하여 찾아낸 지식



통계를 들어가기 전에…

- 통계학의 어원
  - 기원전 435년 로마에서는 시민 등록을 위해 인구조사를 전담하는 '켄소르 (censor)'라는 관리가 있었으며,
  - 오늘 날 여론 조사를 뜻하는 '센서스 (census)'라는 단어는 여기에서 유래 함
  - 통계학 (statistics)은 라틴어의 status (state, 국가)에서 유래되었으며, 원래는 국가산술 (state arithmetic)이란 뜻

통계를 들어가기 전에...

- 일상적인 의사결정 과정과 통계는 뭐가 다를까?
  - 우리가 일상생활에서 무의식적으로 사용하는 방식은 결정론적 의사결정 과정입니다.
    - 새로운 판매전략으로 매출이 1000만원 올랐으니 성공이다.
    - 학원을 다녔더니 수학점수가 10점이나 올랐어
    - 남자친구와 헤어지고 나니 체중이 5kg나 빠졌어.
    - 변화의 원인을 결정론적으로 확신하는 것이 결정론적 의사결정이다.

  - 통계적 의사결정이란?
    - 판매전략의 변화로 매출이 우연히 1000만원 오를 가능성은 얼마일까?
    - 학원을 다녀서 수학점수가 우연히 10점 오를 가능성은 얼마일까?
    - 남자친구와 헤어지고 나서 체중이 우연히 5kg 빠질 가능성은 얼마일까?
    - 통계적 의사결정이란 어떤 사건이 우연히 발생할 확률을 묻는 것으로 시작하는 것을 의미한다.

통계를 들어가기 전에...

- 이번 강의에서 이것 하나만 기억합시다.
  - 통계가 어려운 이유는 통계적으로 생각하는 방법이 낯설기 때문이다.

  - 우리가 통계책을 펴고 공부하기 전에 통계적인 의사결정이란 무엇인지, 그리고 통계적으로 생각한다는 것이 무엇인지 분명히 알아야 한다.

  - “어떤 사건이 우연히 발생할 확률이 얼마일까?

통계를 들어가기 전에...

- 통계 교재의 공통점
  - 모든 통계 교과서의 첫 페이지는 항상 평균과 표준편차로 시작한다.
  - 왜 그럴까?
  - 당연해 보이는 것부터 질문을 해 보아야 한다.
  - 세상에 당연한 것은 없습니다. 여기에도 이유가 있고 꽤나 중요한 이유이다.
  - 통계의 본질은 분산의 마법 이다.

- 머릿속에 아무나 당신과 가장 친한 친구를 머리에 떠올려보자.
- 이 사람을 설명하기 위해 필요한 단어들을 나열해 보자.
  - 키가 작은/ 얼굴이 긴/ 눈이 큰 / 코가 긴/ 입술이 두툼한 ...
- 우리는 자연스럽게 그 사람의 대표적인 특징을 설명한다.
- 그렇다면 당신이 가진 데이터를 설명해 보라고 한다면 우리는 어떻게 해야 할까?

통계를 들어가기 전에...

- 여러분이 가진 데이터의 모습 중 대표적인 특징을 잡아서 설명
- 여기서 말하는 대표적인 특징이라는 것이 데이터의 대표값 이라는 개념
  - 대표값 : 평균, 중간값, 최빈값, 표준편차, 분산, 구간, 최소값, 최대값 ...
- 여기서 평균과 표준편차 (분산)이 등장
- 그런데 왜 다른 대표값들 중에 왜 하필 평균과 표준편차를 더 많이 중요하게 여겨질까?

- 평균
  - 평균 (Mean) = 데이터 전체의 합 (sum) / 데이터의 개수 (n)
  - 평균의 의미 : 데이터의 중심값으로서 데이터의 특성을 대표하는 값
  - 계산이 매우 쉬움
  - 평균은 모든 데이터로부터 영향을 받는다. 다시 말해, 이상한 값의 영향을 심각하게 받음 (outlier에 취약)

통계를 들어가기 전에...

- 분산
  - 분산 (variance) : $$v = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$
    - 분산의 의미는 내가 가진 데이터가 평균값을 중심으로 퍼져 있는 평균적인 거리
    - 수식의 분자 부분은 각 데이터 값에서 평균을 뺀 것. +와 –가 섞여 있어 계산 시 의미가 뭉개짐
      - 이 문제를 해결하기 위해 제곱을 하여 강제로 모든 값을 +로 변형

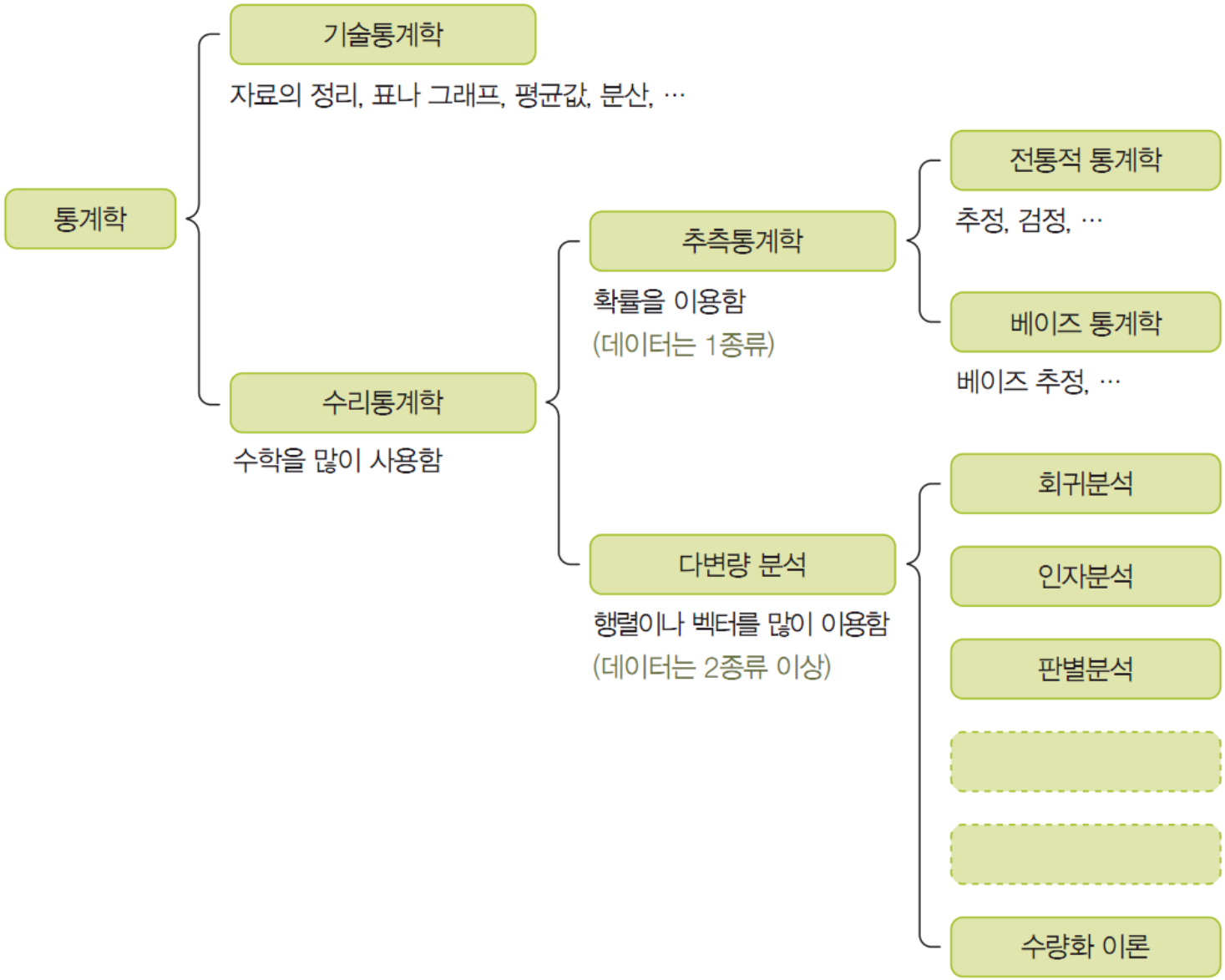
- 표준편차
  - 표준편차 (standard deviation) = root(분산)
  - 의미상으로 표준편차와 분산은 같음

통계를 들어가기 전에...

- 평균과 분산 (표준편차)
  - 데이터의 특성을 표현하는 데 있어 계산이 쉽고 통계를 통해 참값을 추정하기에 가장 효율적
  - 데이터가 많아지면 중간값이나 최빈값은 계산이 아니라 숫자 찾기로 변질
  - 가우스에 따르면 데이터의 불규칙성이 중규분포를 따르고 있으면 최소제곱법이 가장 좋은 추정 방법이고, 그 결과 평균값이 가장 좋은 추정값이 된다고 증명함
  - 체비체프에 따르면 데이터의 불규칙성이 어떠하든 (평균값 +- 2xSD) 범위 안에는 반드시 전체의 ¾ 이상의 데이터가 존재하는 것을 증명함
  - 우리가 공부할 통계는 평균과 분산이 기본이자 중심이며, 그 중에서 분산이 매우 중요함
    - “통계란 분산의 마법 "

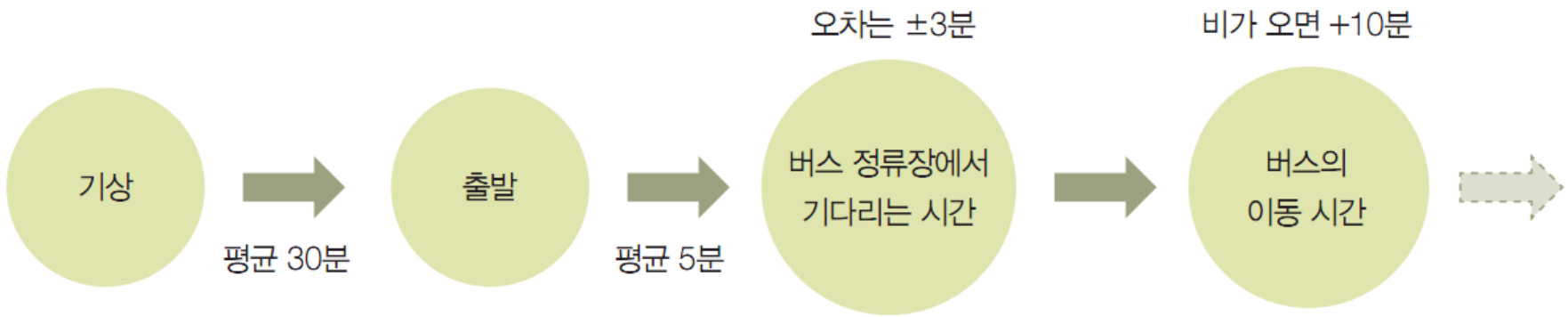
통계를 들어가기 전에...

- 통계학의 분류

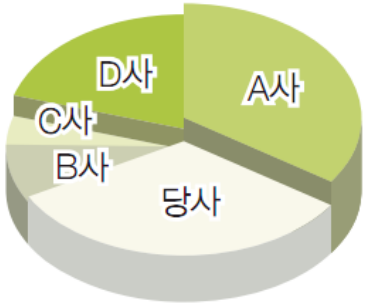


통계를 들어가기 전에...

- 우리 일상생활 자체가 통계
  - 기상과 함께 시작되는 통계적 판단



  - 방대한 통계에 둘러싸인 광고 게시판

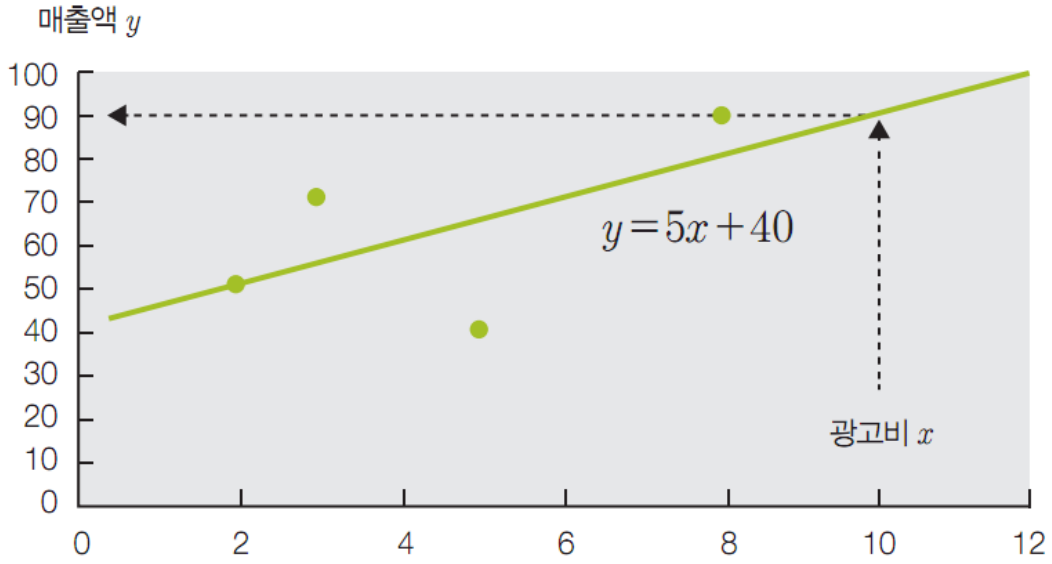


통계를 들어가기 전에...

- 자료를 기초로 통계적 판단 시도하기
  - 2022년 매출액 10억원 목표

연도	광고비 x	매출액 y
2018	2	50
2019	3	70
2020	4	40
2021	8	90

(단위 : 천만 원)



$100=5x+40$에 의해 $x=12$

1억 2000만 원의 광고비가 타당

통계학의 기초 – 자료 정리

- 도수분포표
  - 데이터가 취하는 값의 범위를 몇 개의 구간으로 나누어 데이터가 해당 구간에 몇 개씩 들어가는지 정리
  - 계급이 너무 작으면: 전체의 특징을 파악하기 곤란
  - 계급이 너무 크면: 각 데이터의 차이를 파악하기 곤란

계급 (이상 ~ 미만)	계급값	도수
500 ~ 10000	7500	0
10000 ~ 15000	12500	2
15000 ~ 20000	17500	6
20000 ~ 25000	22500	19
25000 ~ 30000	27500	25
30000 ~ 35000	32500	18
35000 ~ 40000	37500	40
40000 ~ 45000	42500	20
45000 ~ 50000	47500	18
50000 ~ 55000	52500	9
55000 ~ 60000	57500	3
	합계	160

통계학의 기초 – 자료 정리

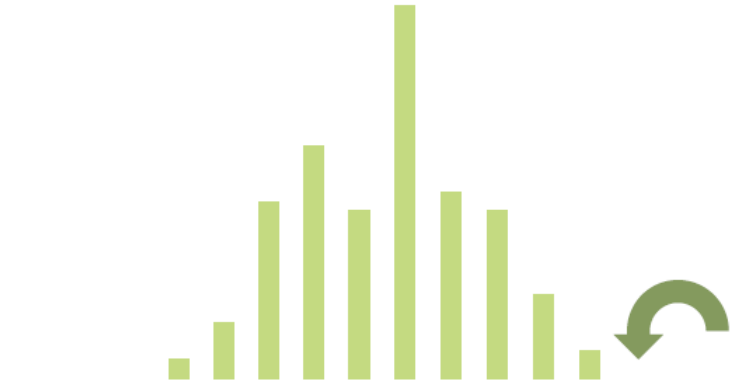
- 상대도수분포표
  - 상대 도수는 각 도수를 총 도수로 나눈 실숫값

계급 이상	계급 미만	계급값	도수	상대도수
500	10000	7500	0	0.00000
10000	15000	12500	2	0.01250
15000	20000	17500	6	0.03750
20000	25000	22500	19	0.11875
25000	30000	27500	25	0.15625
30000	35000	32500	18	0.11250
35000	40000	37500	40	0.25000
40000	45000	42500	20	0.12500
45000	50000	47500	18	0.11250
50000	55000	52500	9	0.05625
55000	60000	57500	3	0.01875
		합계	160	1

25/160

통계학의 기초 – 자료 정리

- 도수분포그래프
  - 시계 반대 방향으로 90도 회전한 도수를 높이로 하는 직사각형이 자료의 특징을 더 확실히 파악 가능

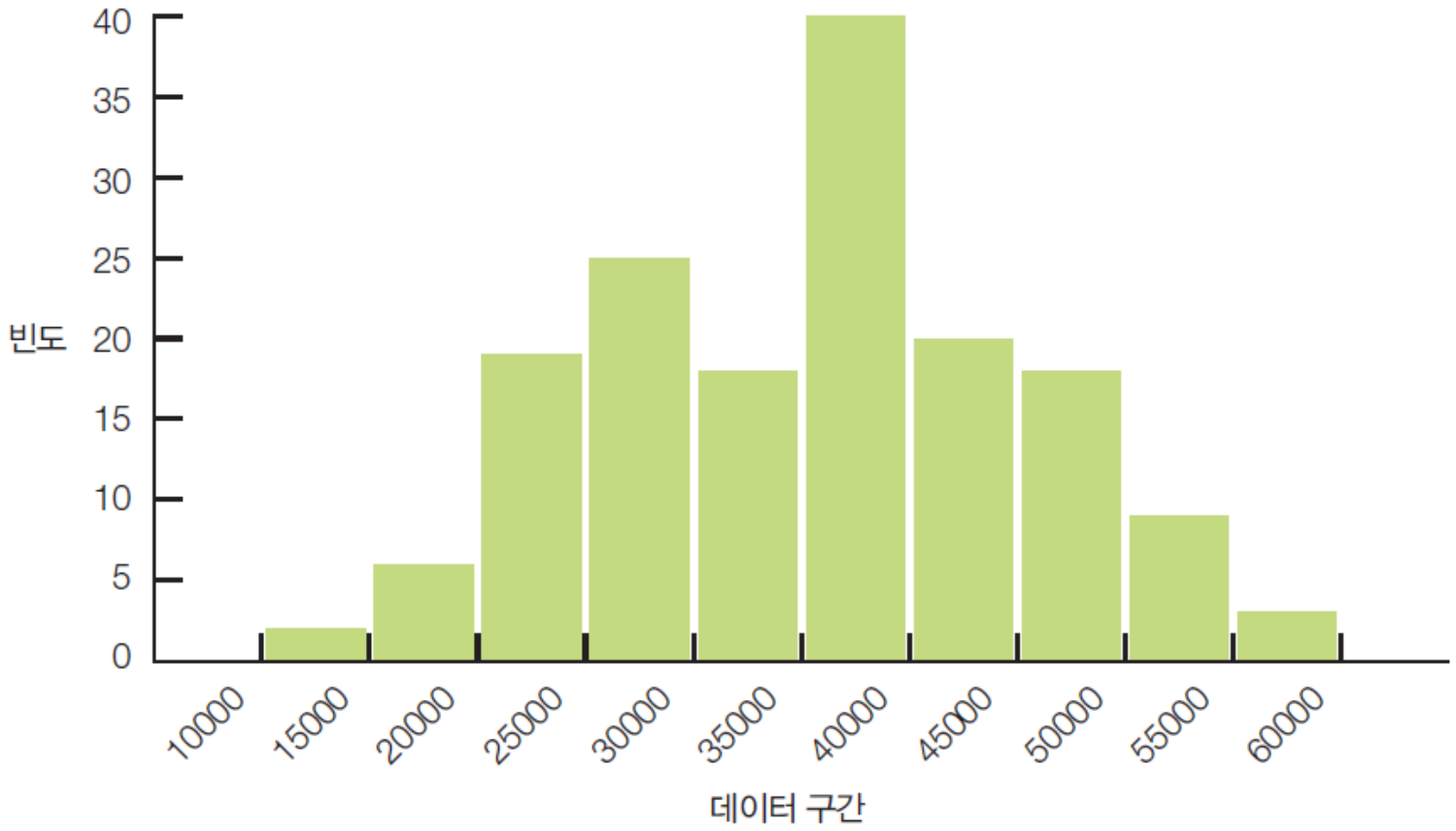


계급 (이상 ~ 미만)	계급값	도수
500 ~ 10000	7500	0
10000 ~ 15000	12500	2
15000 ~ 20000	17500	6
20000 ~ 25000	22500	19
25000 ~ 30000	27500	25
30000 ~ 35000	32500	18
35000 ~ 40000	37500	40
40000 ~ 45000	42500	20
45000 ~ 50000	47500	18
50000 ~ 55000	52500	9
55000 ~ 60000	57500	3
	합계	160

계급 (이상 ~ 미만)	계급값	도수
500 ~ 10000	7500	0
10000 ~ 15000	12500	2
15000 ~ 20000	17500	6
20000 ~ 25000	22500	19
25000 ~ 30000	27500	25
30000 ~ 35000	32500	18
35000 ~ 40000	37500	40
40000 ~ 45000	42500	20
45000 ~ 50000	47500	18
50000 ~ 55000	52500	9
55000 ~ 60000	57500	3
	합계	160

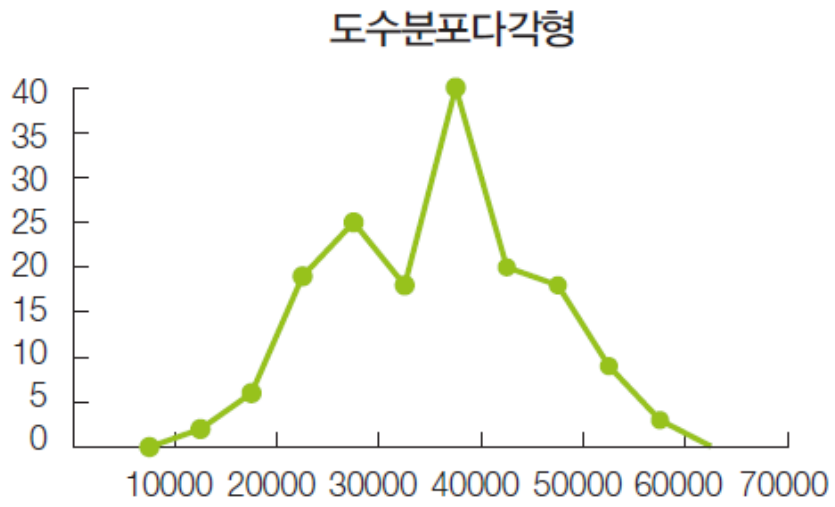
통계학의 기초 – 자료 정리

- 히스토그램
  - 도수분포그래프의 막대 사이에 있는 간격을 없앤 그래프



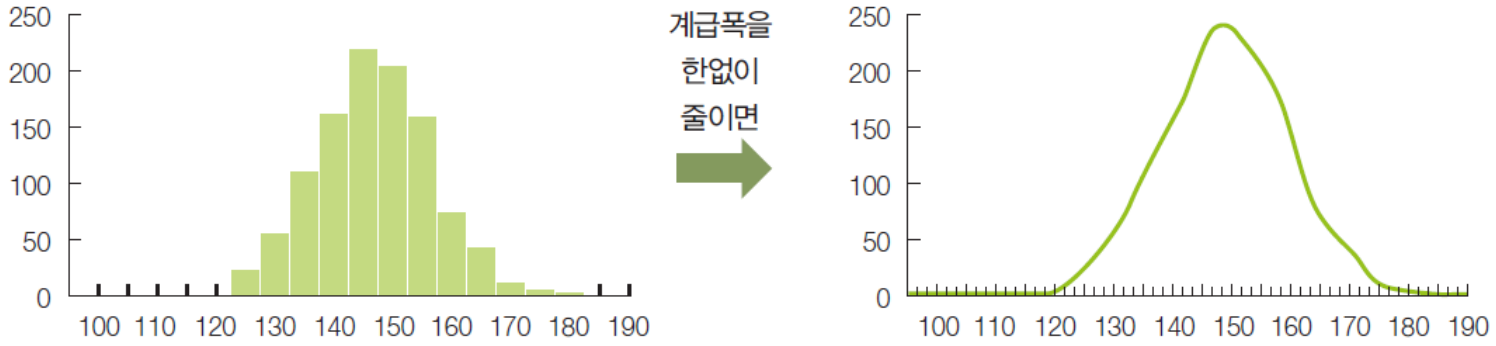
통계학의 기초 – 자료 정리

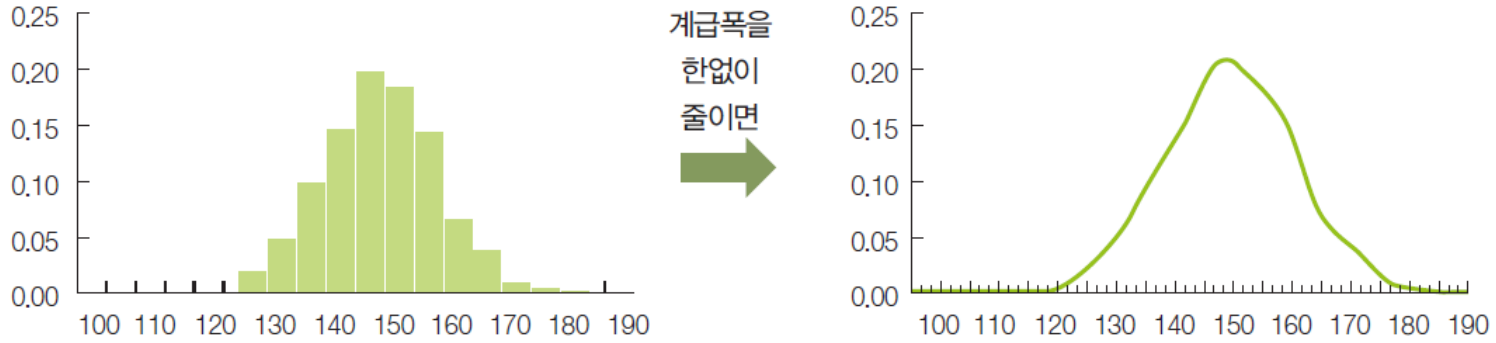
- 도수분포다각형
  - 히스토그램의 기둥 윗부분의 중점을 차례로 연결해서 생기는 꺾은선 그래프



통계학의 기초 – 자료 정리

- 도수분포곡선
  - 이산 변량의 경우 계급폭(구간)을 줄이는데 한계가 있지만,
  - 연속 변량의 경우 구간폭을 줄이면 히스토그램은 곡선이 됨





통계학

- 수치 데이터의 수집, 분석, 해석, 표현 등을 다루는 수학의 한 분야로 크게 기술 통계학과 추론 통계학으로 분류

- 기술 통계 (descriptive statistics)
  - 수집한 데이터를 요약, 묘사, 설명하는 통계 기법
  - 변수 형태 표현: 표, 그래프
  - 변수의 특성값 도출 : 평균, 분산, 표준편차 …
- 추론 통계 (inferential statistics)
  - 수집한 데이터를 바탕으로 추론 예측하는 통계 기법
  - 추리통계, 추측통계 라고도 불리며, 기술통계가 선행되어야 함
  - 모집단의 통계적 특성을 구하기 위해 표본 (sample)의 통계적 특성 (기술 통계)을 통해 추측
    - 실제 모집단의 통계적 특성과 차이가 있을 수 있으며, 이를 표집오차 (sampling error)라고 함
    - 따라서 표본의 통계적 특성이 해당 표본에 한정된 현상인지, 아닌 지 검증해야 함

- 통계 모델링 : 데이터에 통계학을 적용해 변수의 유의성을 분석함으로써 데이터의 숨겨진 특징을 찾아내는 것

변수

- 변수 (variable)
  - 하나의 개념을 대표하는 상징으로서 그것의 특성이 갖는 값이나 강도 또는 크기의 차이를 나타낼 수 있는 것
  - 값이 달라질 수 있는 것
  - 독립변수 (independent variable) : 원인이 되는 것 (설명변수)
  - 종속변수 (dependent variable) : 결과가 되는 것 (결과변수 / 반응변수)
  - 예시 : 출생 시 체중은 병원 감염 발생에 영향을 주는가?
    - 독립변수 : 출생 시 체중
    - 종속변수 : 병원감염 발생

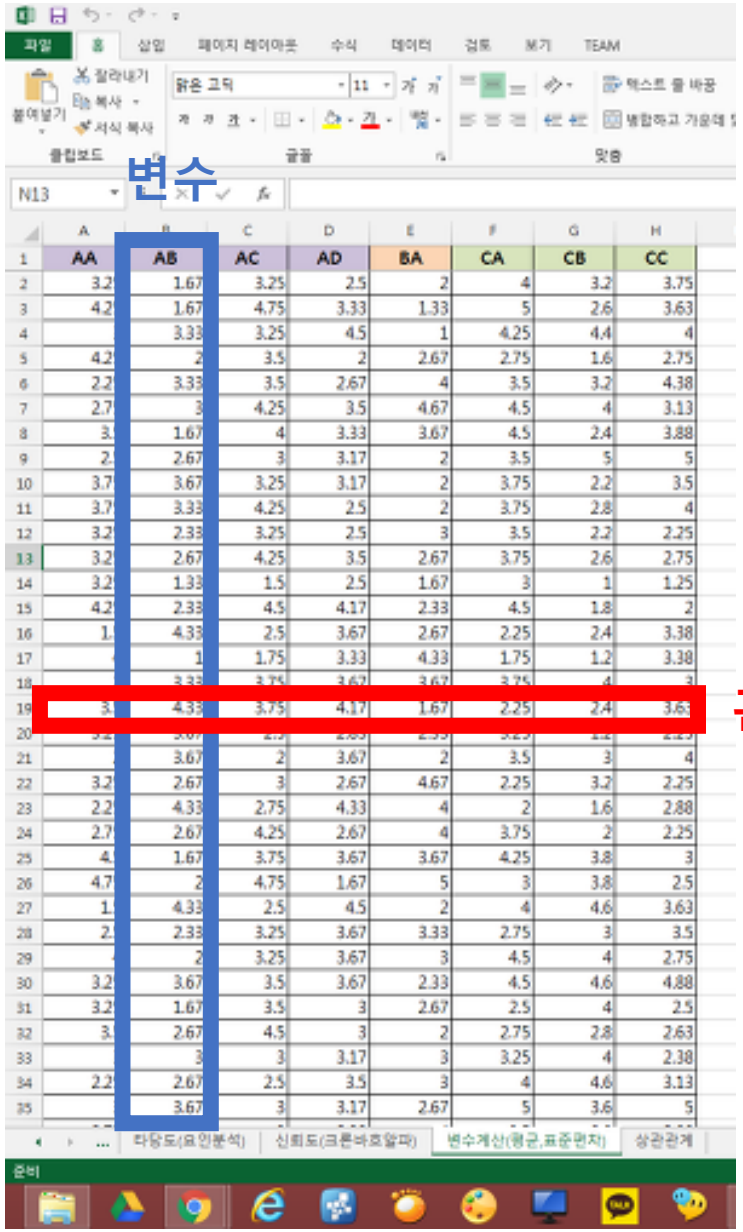
- 상수 : 변하지 않고 항상 그대로 있는 숫자
  - pi = 3.141592 …

변수

워크샵의견 설문조사지

■ 의견

설문내용	평가				
	매우 적절	적절	그저 그렇다	부적절	매우 부적절
1. 금번 실시한 워크숍의 일정 / 시간계획의 적절성은?					
2. 워크샵 장소와 프로그램은 적절하였는가?					
3. 워크샵 기간 동안의 숙소 / 식사는 적절하였는가?					
4. 금번 워크샵 진행이 원활하고 매끄럽게 이루어졌는가?					
5. 금번 워크샵을 통한 직원 결속력 강화 및 커뮤니케이션활성화에 끼친 영향을 평가한다면?					
6. 금번 워크샵이 전반적으로 만족도 하였는가?					



변수


번호
이름
성별
나이
신장
건강 상태
1
김영대
남
35
172.5
양호
2
박혜은
여
28
168.3
보통
3
…
…
…
…
…
4
…
…
…
…
…
…
…
…
…
개체명
변량
(변수)
개체
변량 값
(데이터)

기술 통계 – 변수

- 변수의 종류 : 변수는 크게 질적 변수 (qualitative variable)와 양적 변수 (quantitative variable)로 분류

- 질적 변수 : 변수의 크기를 정의할 수 없는 형태의 범주형 변수
  - 보기의 순서가 바뀌어도 상관 없음
  - 평균, 표준편차 등을 사용할 수 없으며, 덧셈, 뺄셈과 같은 사칙연산도 사용할 수 없음
  - 대신 응답자 수와 백분율 중요
  - 예시 : 성별, 인종, 혈액형 등

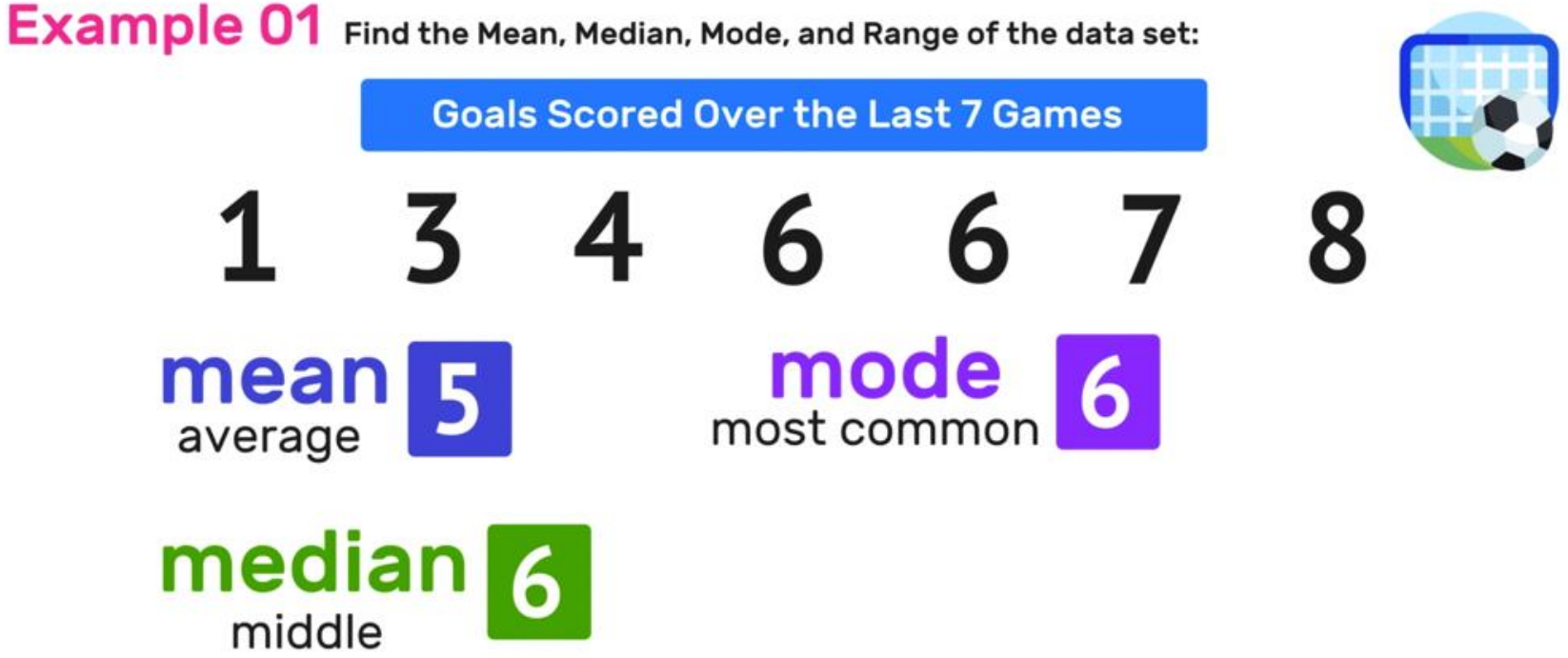
- 양적 변수 : 변수의 크기와 순서를 정의할 수 있는 형태의 수치 변수
  - 연속형 변수 (continuous variable)
    - 예시: 몸무게, 키, 온도, 수익률, 시장 점유율
  - 이산형 변수 (discrete variable)
    - 예시: 서울시 아파트 층수, 퀴즈를 맞춘 개수, 복용한 알약 개수
  - 평균과 표준편차와 같은 기술 통계 값이 중요

기술 통계 - 변수

- 다음의 변수는 질적변수 일까 양적변수 일까?
  - 색상
  - 무게
  - 가격
  - 둘레 (지름)
  - 브랜드
  - 모델번호
  - 부채비율
  - 소프트웨어 버전
  - 자동차 년식 (년도)
  - 속도 (km/h)
  - 온도
  - 여행예산
  - 스카우트 순위

기술 통계 – 양적 변수

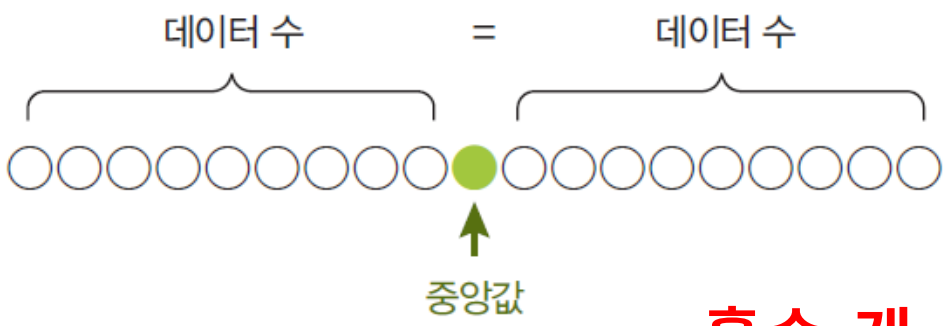
- 기술 통계는 크게 중심화 경향과 분산도로 구분
- 중심화 경향 (Central tendency) : 수집한 변수 전체를 대표하는 값이 무엇인지 나타내는 통계 (대표값)
  - (산술) 평균값 [mean] : 변수를 모두 더해서 전체 변수의 개수로 나눈 값
  - 중앙값 [median] : 변수를 크기 순으로 정렬했을 때, 중앙에 위치하는 값
  - 최빈값 [mode] : 변수 중 그 빈도가 가장 많이 나타내는 값



https://www.mashupmath.com/blog/mean-median-mode-range-guide

기술 통계 – 양적 변수

- 중앙값 [median] : 변수를 크기 순으로 정렬했을 때, 중앙에 위치하는 값

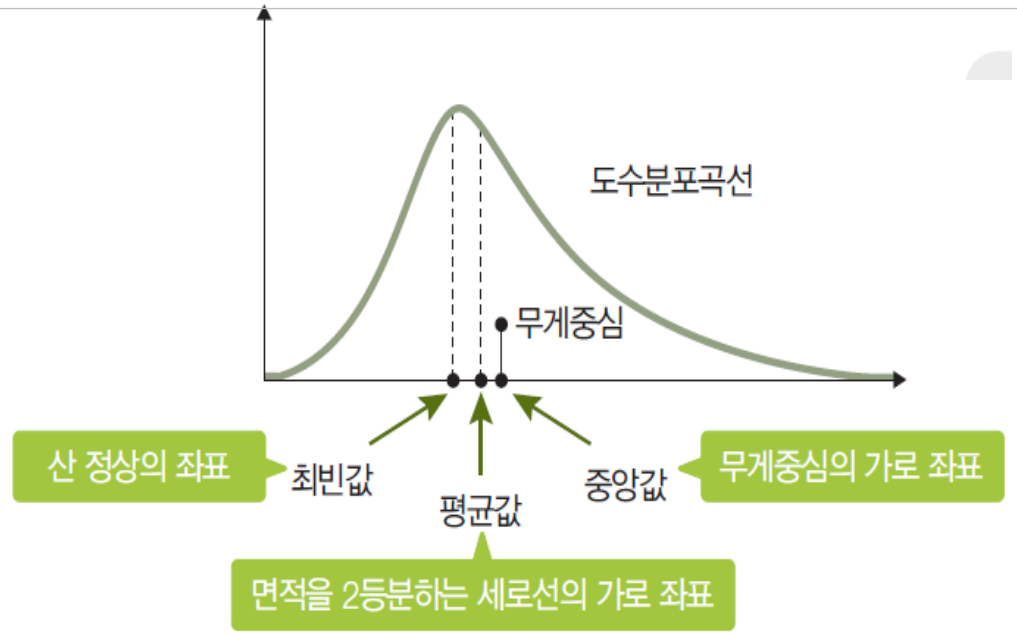


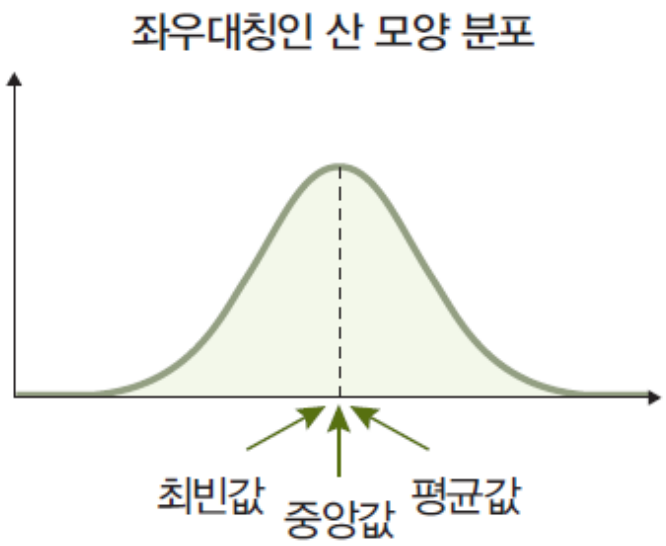
홀수 개

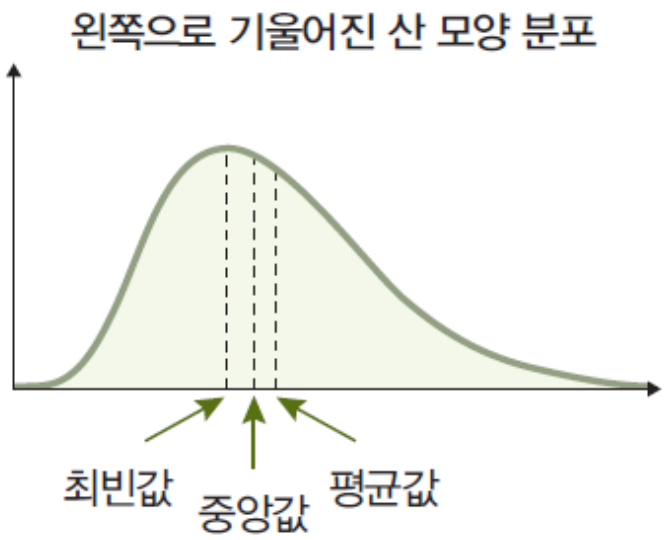
짝수 개

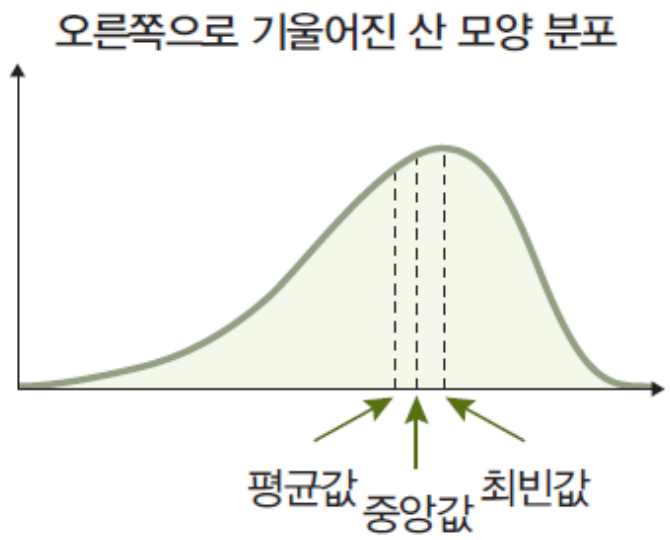
기술 통계 – 양적 변수

- 도수분포곡선









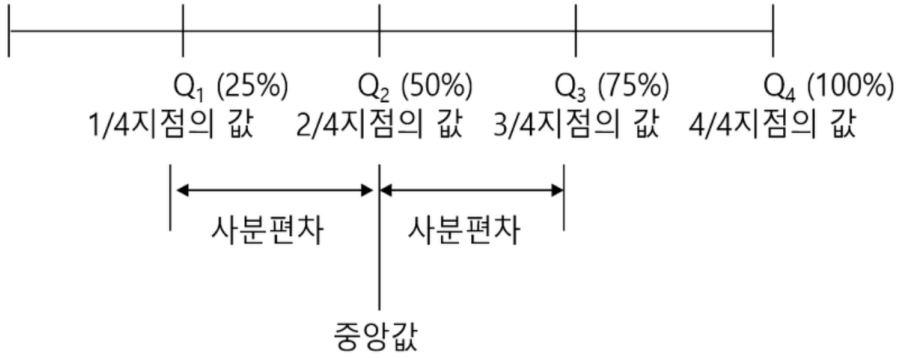
기술 통계 – 양적 변수

- 평균값, 중앙값, 최빈값 만으로는 분포의 특징을 말할 수 없음
  - 두 자료의 평균값은 같아도 내용은 다름

신장의 평균은 같아 보이지만 고르기는 다르다

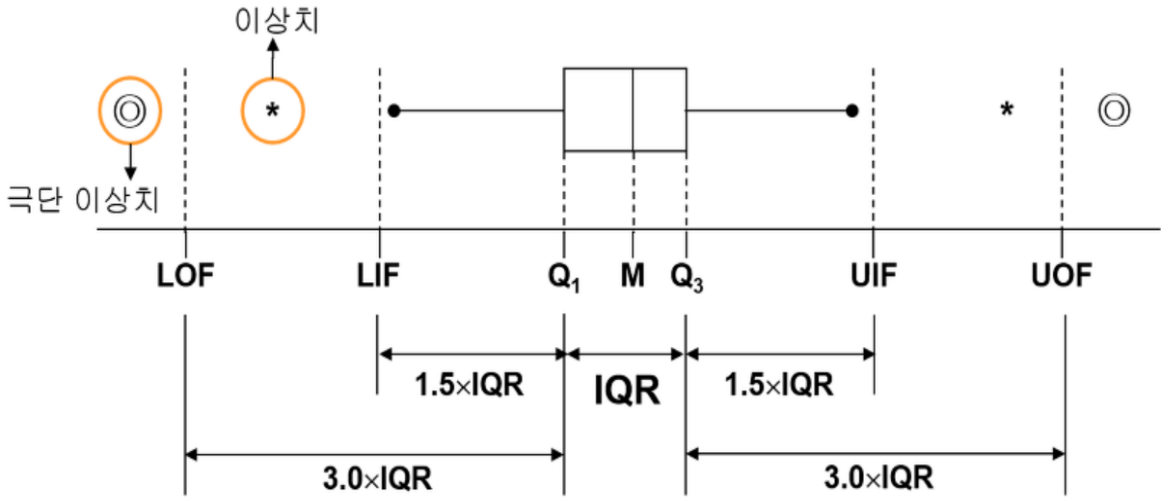
기술 통계 – 양적 변수

- 기술 통계는 크게 중심화 경향과 분산도로 구분
- 분산도 (Variation) : 데이터가 어떻게 분포되어 있는 지를 설명하는 통계치 (산포도)
  - 범위 (range) : 변수의 최대값에서 최소값의 차이
  - 사분편차 (quatile deviation) : 전체 데이터의 ¼, ¾ 지점 사이의 데이터 분포를 보는 통계치
    - 박스 플롯 (box plot)과 연관성



$$Quartile\ deviation = \frac{Q_3 - Q_1}{2}$$

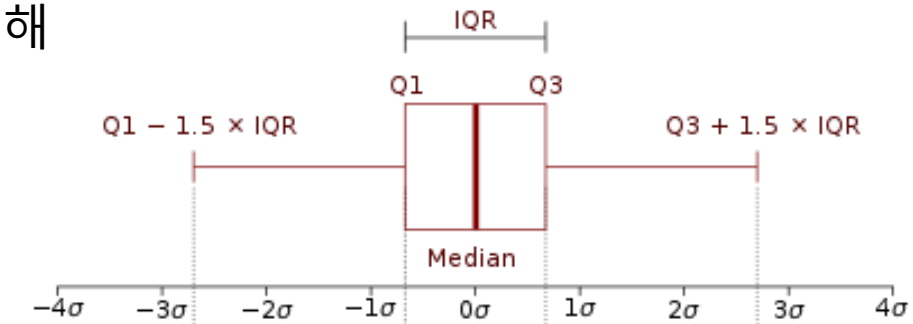
사분편차



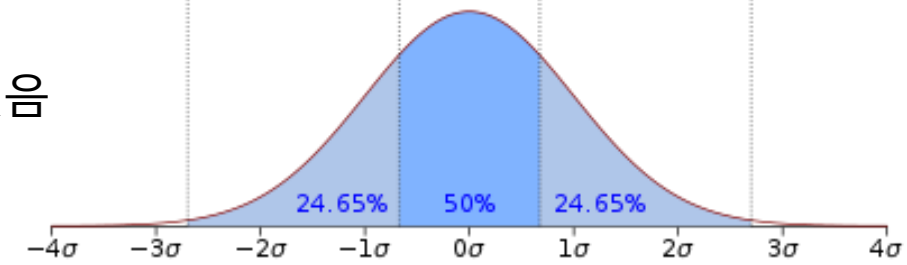
박스플롯

기술 통계 – 양적 변수

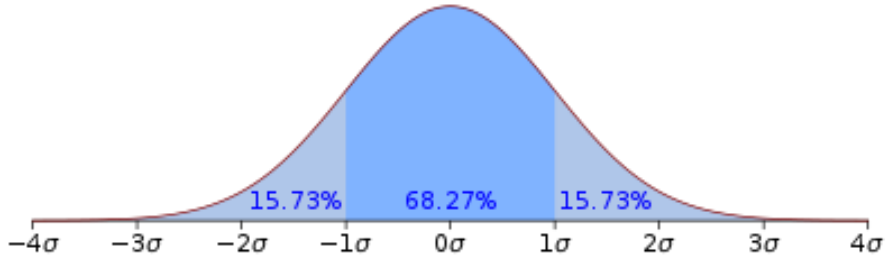
- 기술 통계는 크게 중심화 경향과 분산도로 구분
- 분산도 (Variation) : 데이터가 어떻게 분포되어 있는 지를 설명하는 통계치 (산포도)
  - 분산 (variation) : 개별값과 평균값의 차이를 제곱한 후 다 더해서 전체 변수의 개수로 나눈 값
    - 제곱을 하는 이유는 편차의 합이 0이 되는 것을 방지하기 위해
  - 표준 편차 (standard deviation) : 분산에 제곱근을 취한 값



정규 분포와 비교했을 때 분포를 비교할 수 있음



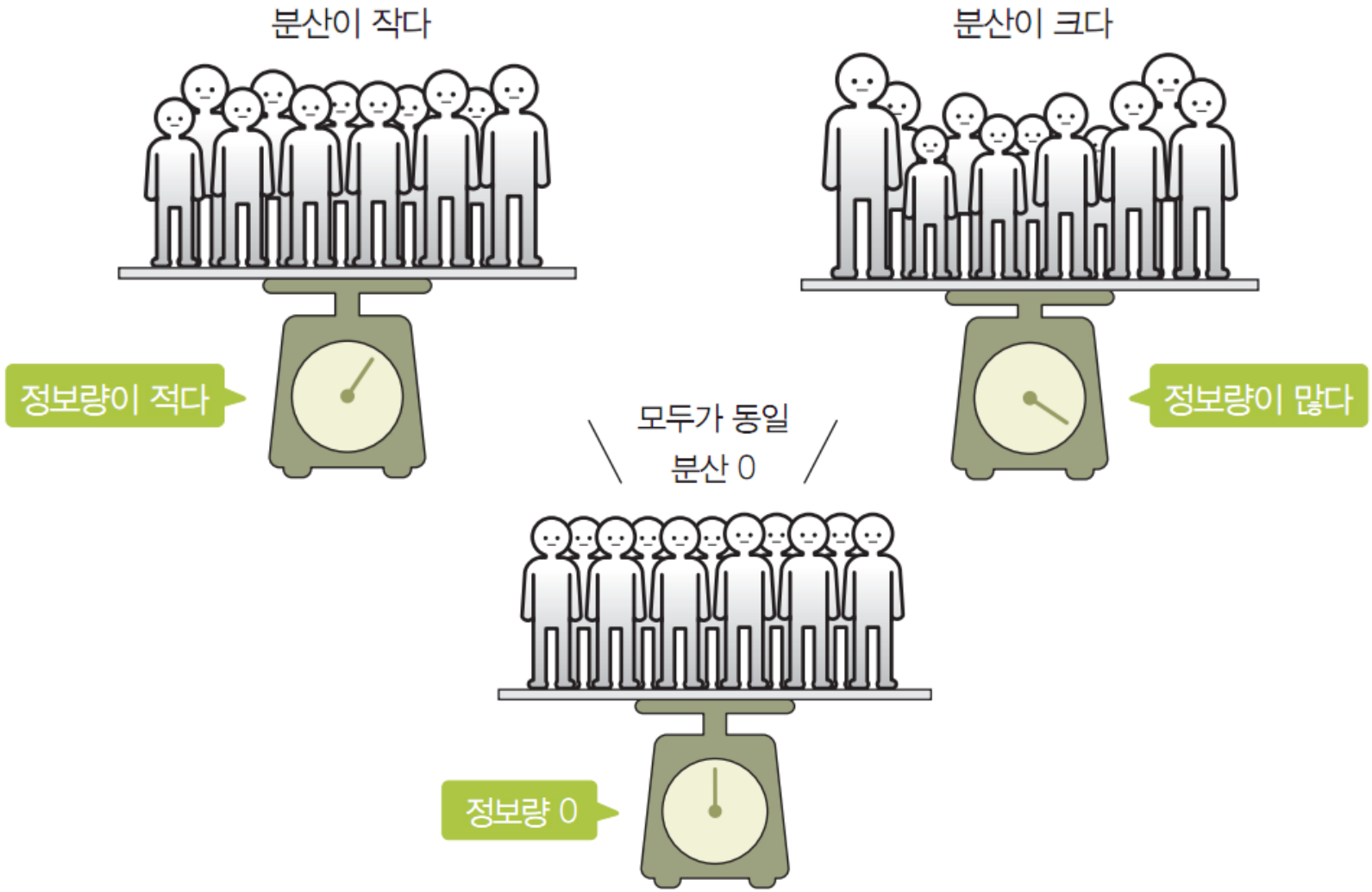
정규분포를 따르면 좌우 1 표준편차 사이에
전체 데이터의 68.27% 가 분포



https://zetawiki.com/wiki/%EC%82%AC%EB%B6%84%EC%9C%84%EB%B2%94%EC%9C%84_IQR

기술 통계 – 양적 변수

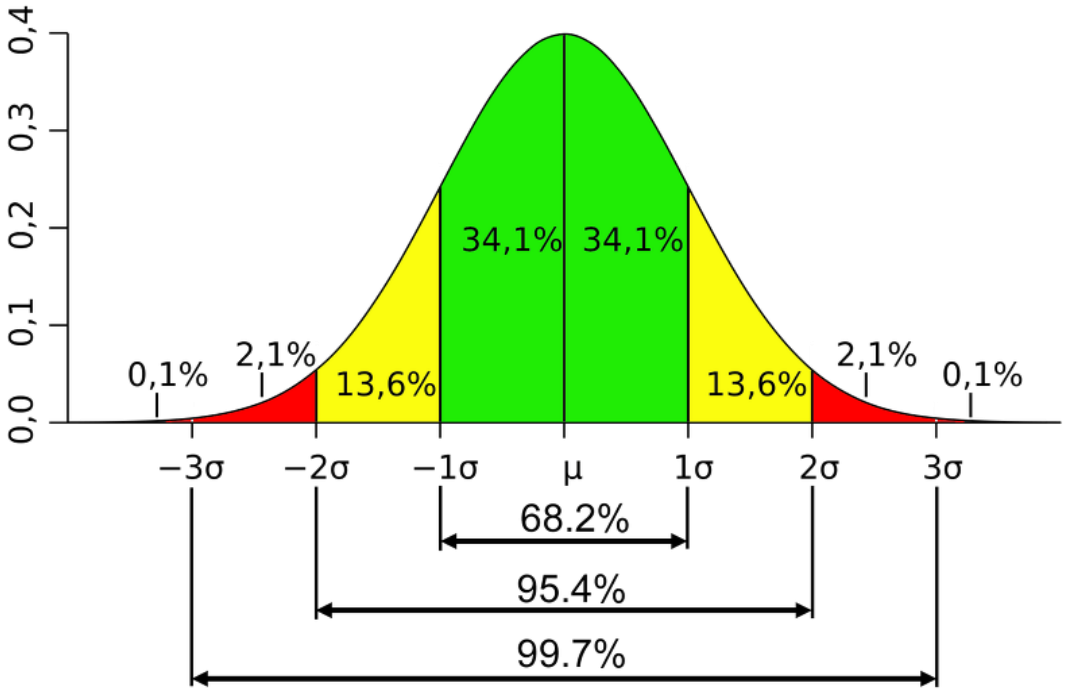
- 분산 (variation)은 통계학의 생명
  - 분산은 데이터의 정보량, 즉 데이터의 다양성 정도를 표현



  - 분산 = 0 ?

정규 분포

- 정규 분포의 특징
  - 평균값 μ를 중심으로 좌우 대칭
  - 곡선은 μ의 근처에서 높고 양측으로 갈수록 낮아짐
  - 표준편차 σ가 곡선의 모양을 결정
    - σ 값이 크면 곡선은 평평해지고, σ 값이 작으면 좁고 높아짐
  - 정규 분포의 어느 구간을 취할 때 그 속에 포함된 전체에 대한 비율을 그 면적의 크기로 알 수 있음



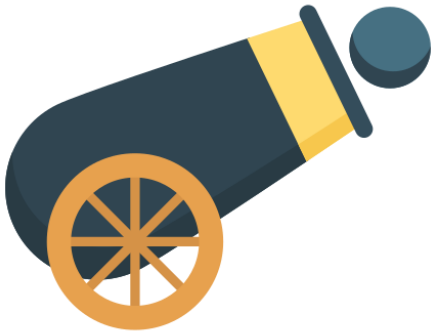
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

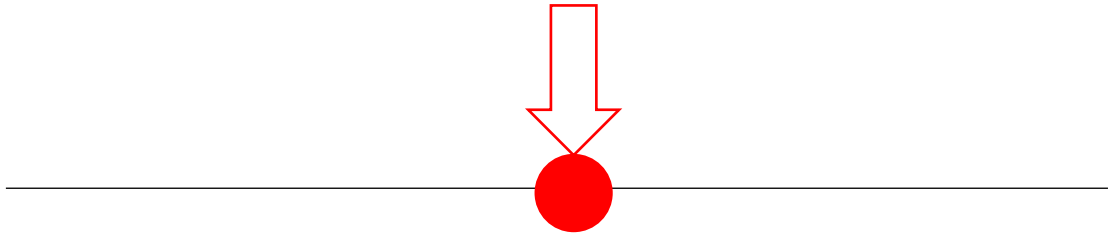
μ : mean (평균)

σ : standard deviation (표준편차)

정규 분포

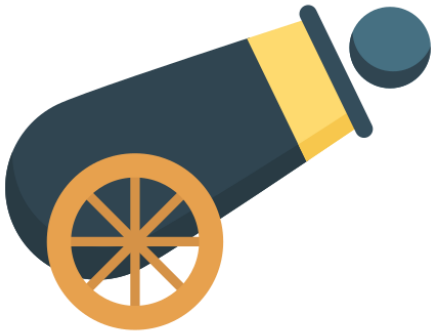
- 오차 분포

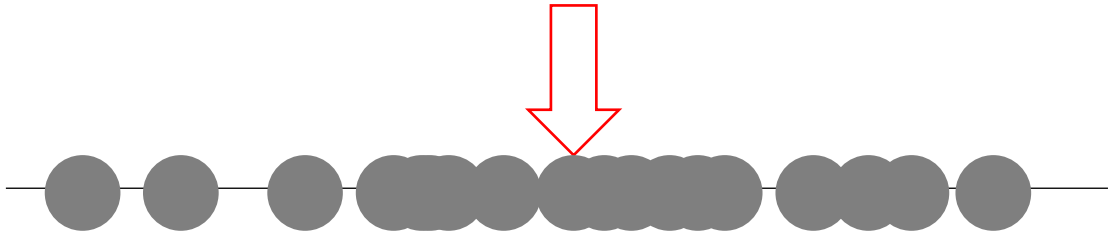




정규 분포

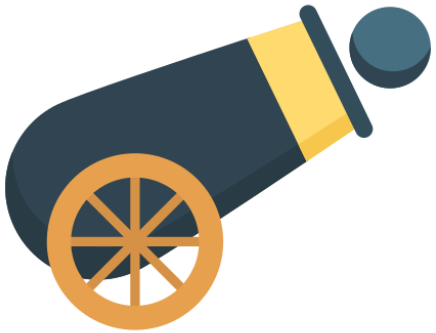
- 오차 분포

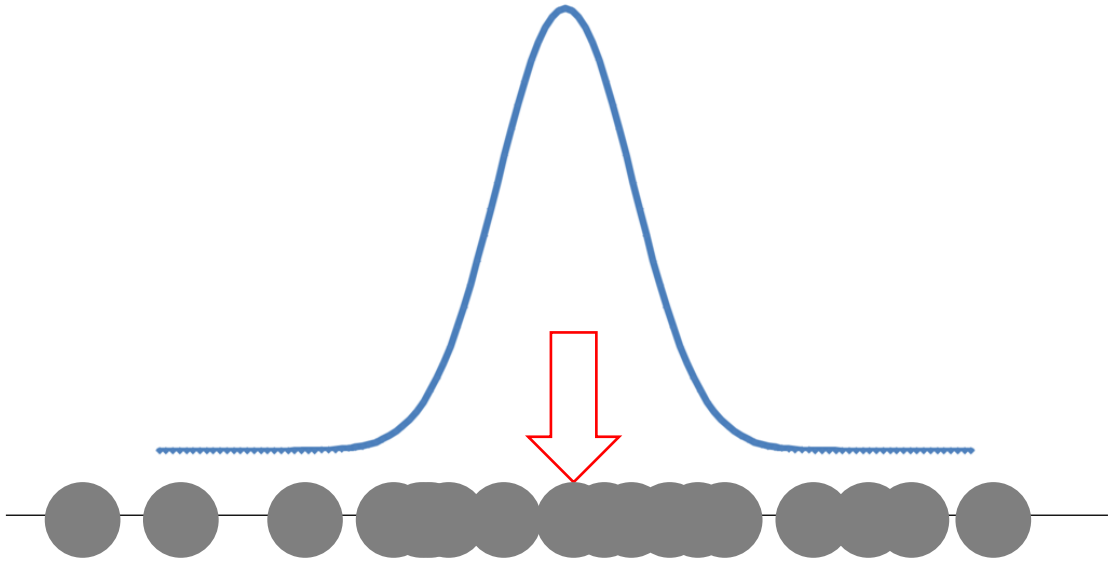




정규 분포

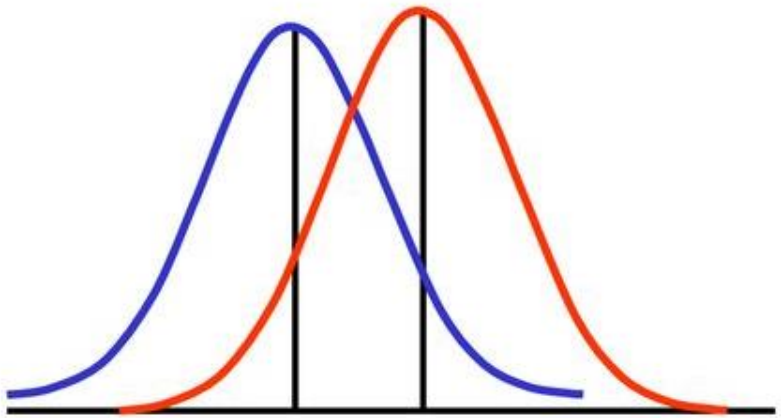
- 오차 분포





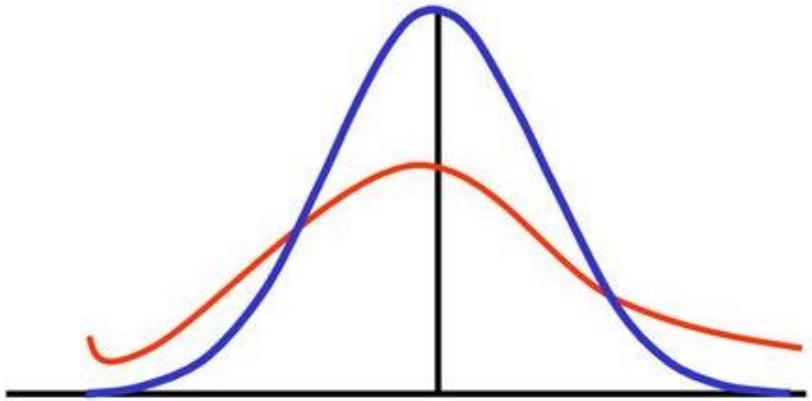
정규 분포

평균은 다르고 분산이 같은 경우



분산이 같으면 분포의 모양이 동일합니다.

분산은 다르고 평균이 같은 경우

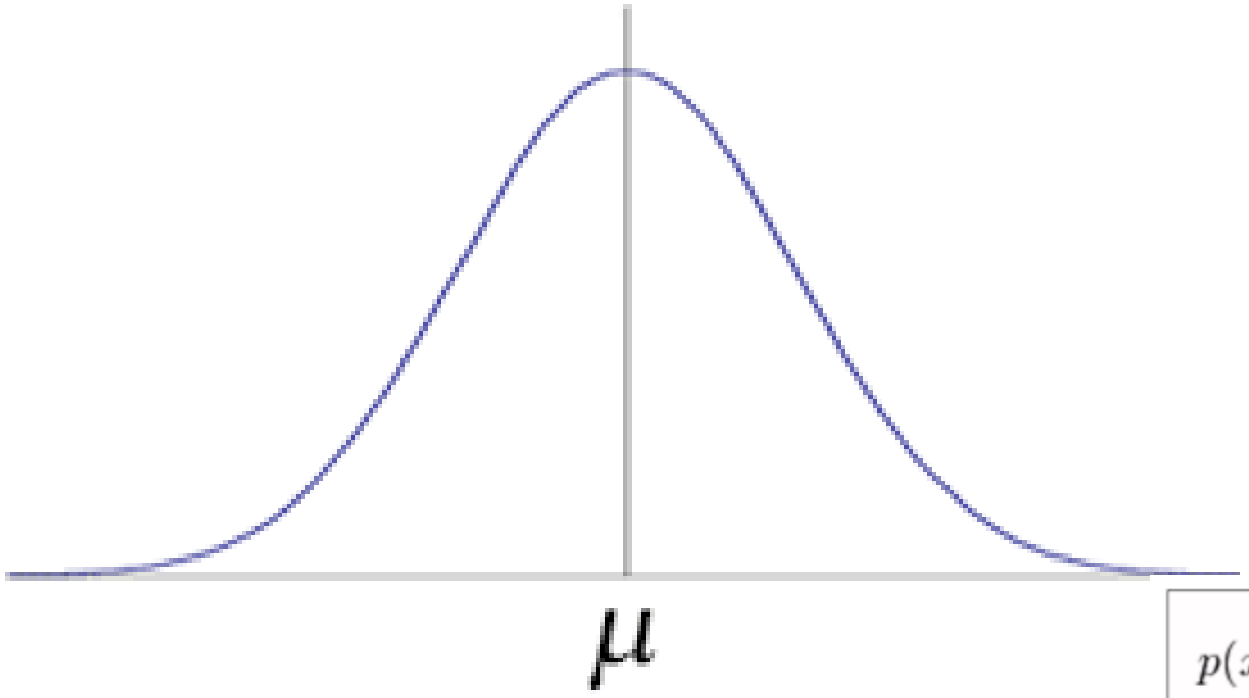


분산이 다르면 분포의 모양이 다르며 분산이 클수록 종모양의 형태가 옆으로 더 펴지는 형태로 나타납니다.

표준 정규 분포

- 면적을 구하려면 적분 계산 : 정규 분포의 모양이 변하기 때문에 적분이 어려움

우리 회사에서 180cm 이하의 직원은 얼마나 있을까?



μ = 175

σ = 5

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ : mean (평균)

σ : standard deviation (표준편차)

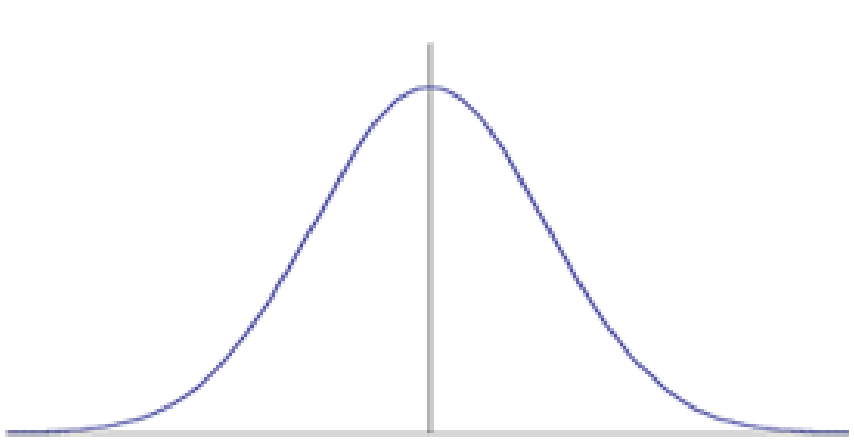
표준 정규 분포

- 모든 정규분포를 평균을 0, 표준편차를 1이 되도록 표준화한 것으로, 어떠한 관측값 x의 값이 그 분포의 평균으로부터 표준편차의 몇 배 정도나 유리되어 있는 가를 표준화된 정규분포 확률변수 z로 나타냄 (z 분포)

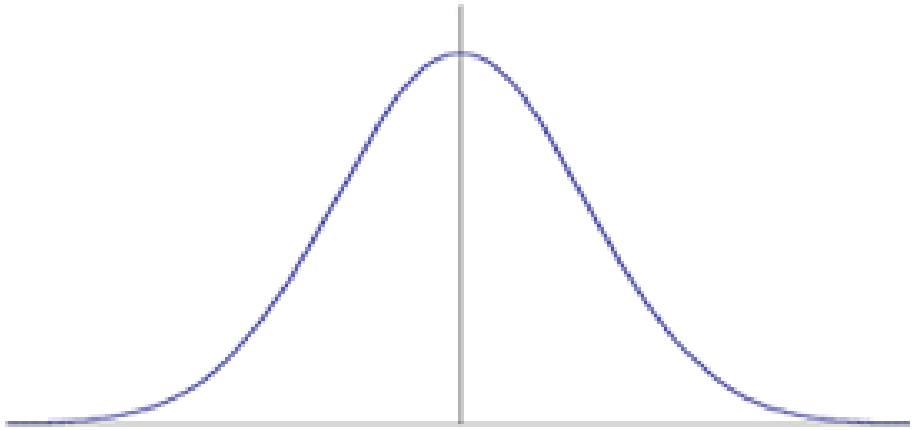
$$Z = \frac{X - \mu}{\sigma}$$

표준 정규 분포

우리 회사에서 180cm 이하의 직원은 얼마나 있을까?



μ = 175
σ = 5

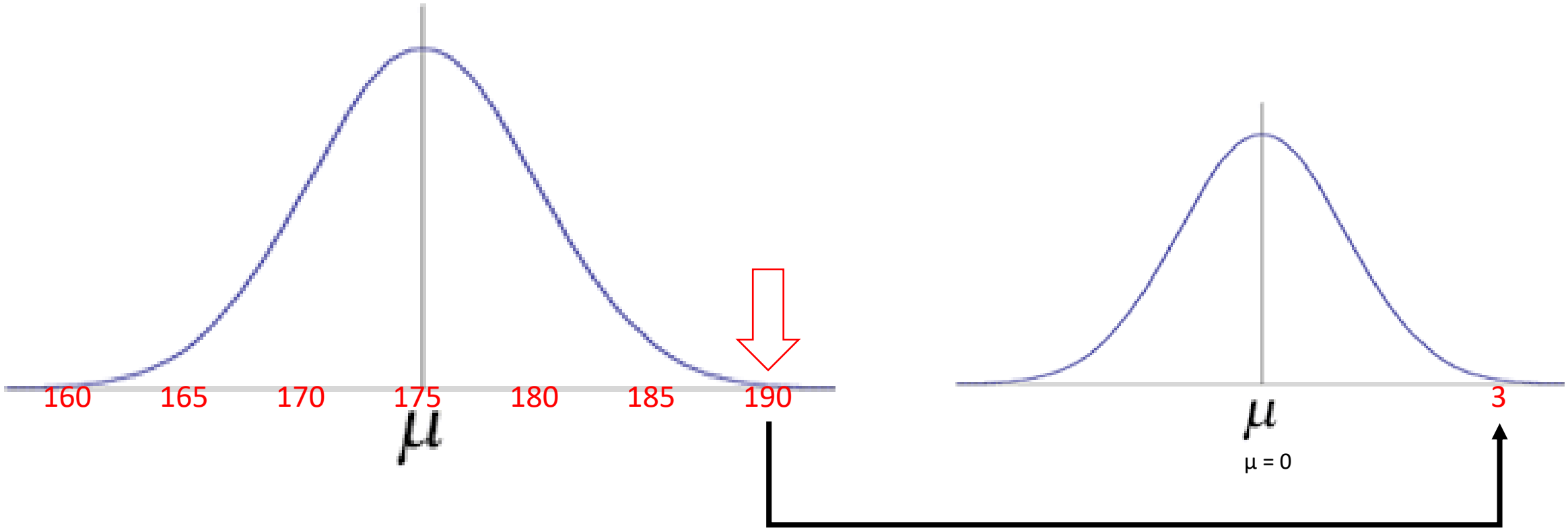


μ = 0

$$Z = \frac{190 - 175}{5} = 3$$

표준 정규 분포

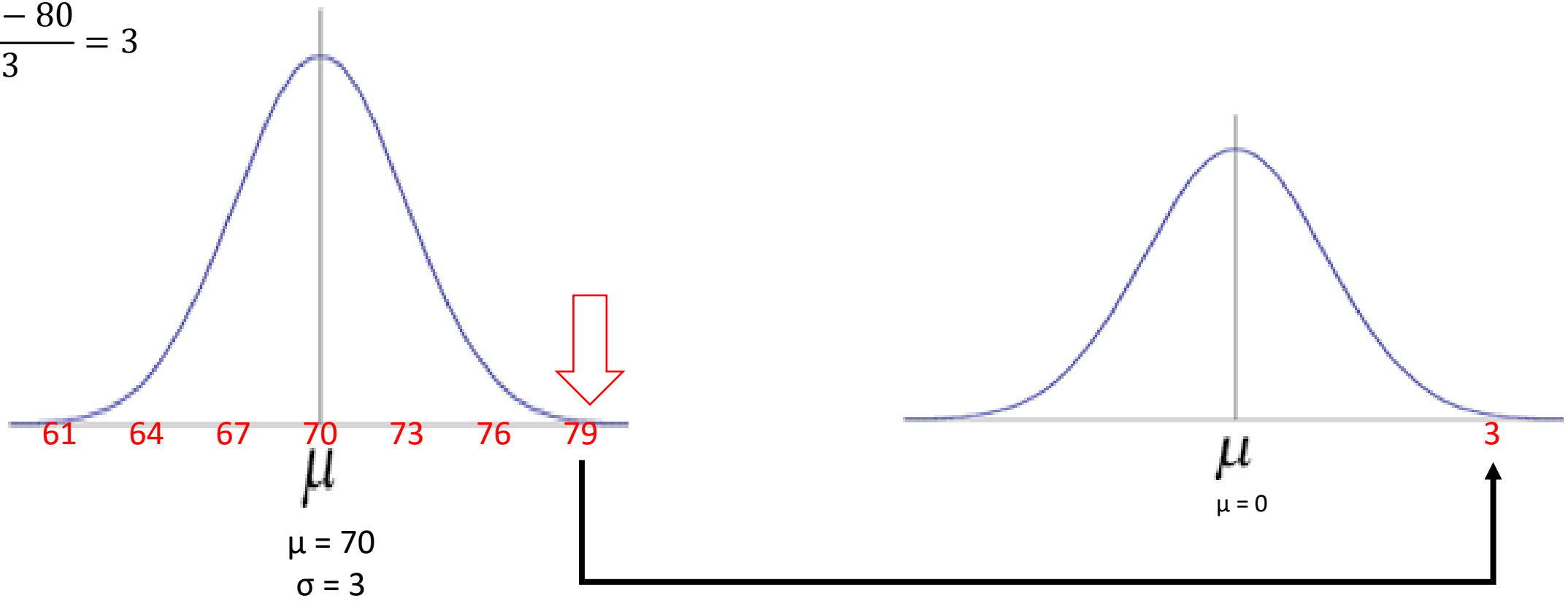
$$Z = \frac{190 - 175}{5} = 3$$



표준 정규 분포

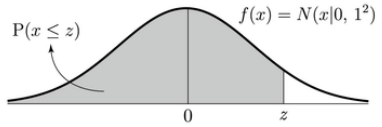
- 정규분포의 표준화 → 표준 정규 분포

$$Z = \frac{79 - 80}{3} = 3$$



표준 정규 분포표

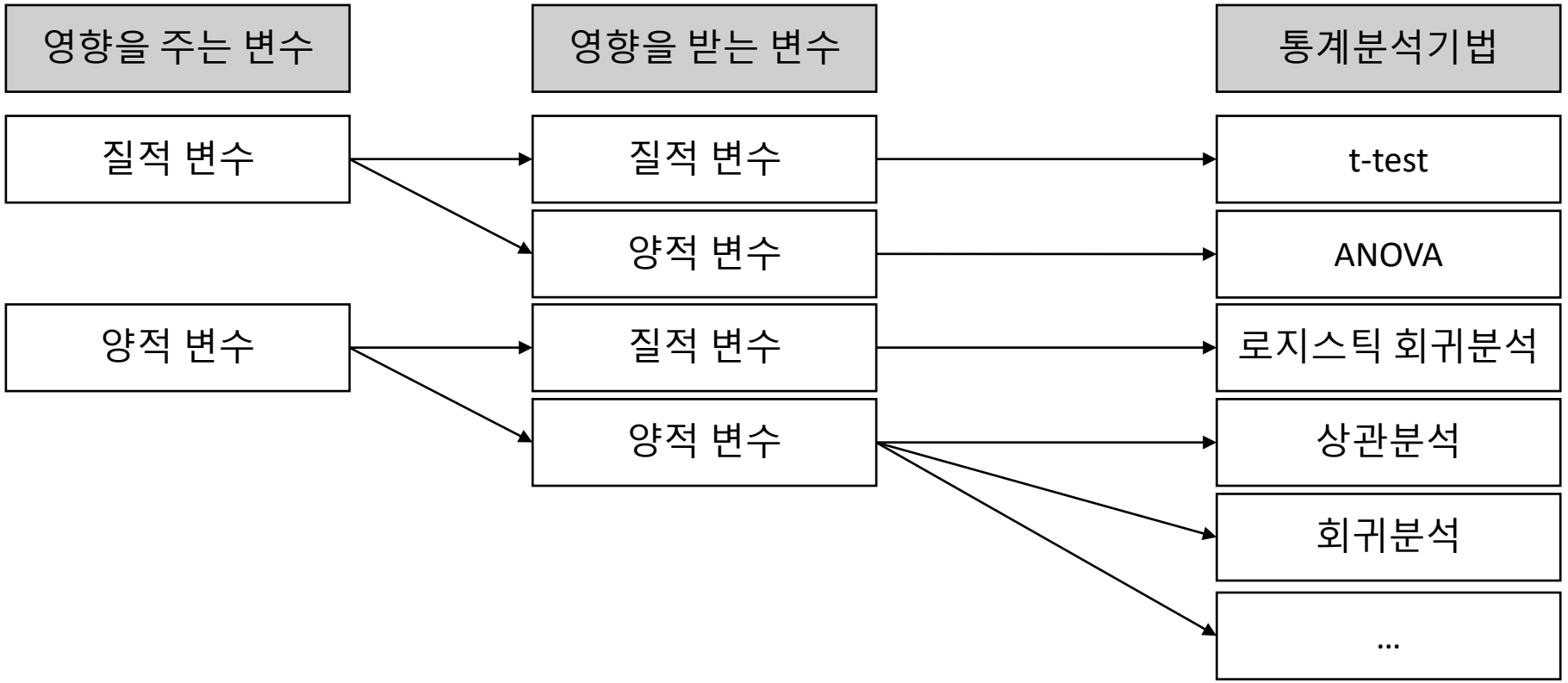
Standard Normal Distribution table



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999							

기술 통계 - 변수

- 변수 (변수)를 구분하는 이유는 서로 다른 통계분석 기법 때문



추론 통계

- 추론 통계는 추정 (estimation)과 가설검정 (testing hypothesis)로 나눌 수 있음

- 추정 (estimation) : 표본을 통해 모집단 특성이 어떠한가 에 대해 추축하는 과정
  - 표본 평균 계산을 통해 모집단 평균을 추측해보거나, 모집단 평균에 대한 95% 신뢰구간의 계산

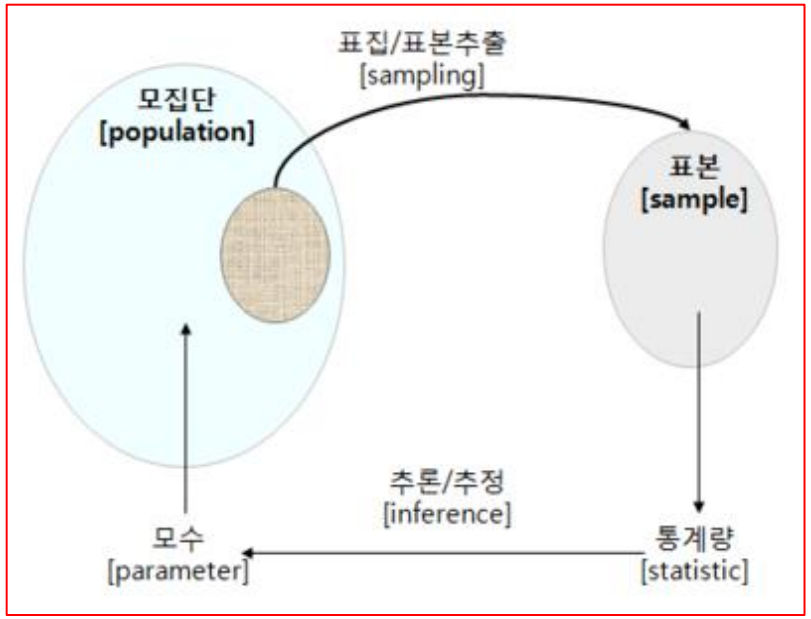
- 가설검정 (testing hypothesis) : 모집단 실제 값이 얼마나 되는가 하는 주장과 관련해서, 표본이 갖고 있는 정보를 이용해 가설이 정확한 지 판정하는 과정

추론 통계 – 모집단과 표본

- 모집단과 표본
  - 모집단 (population distribution) : 통계적인 실험이 되는 모든 대상들의 집합
  - 모수 (parameter) : 모집단의 특성을 나타내는 수치
    - 모평균, 모분산, 모표준편차, 모비율

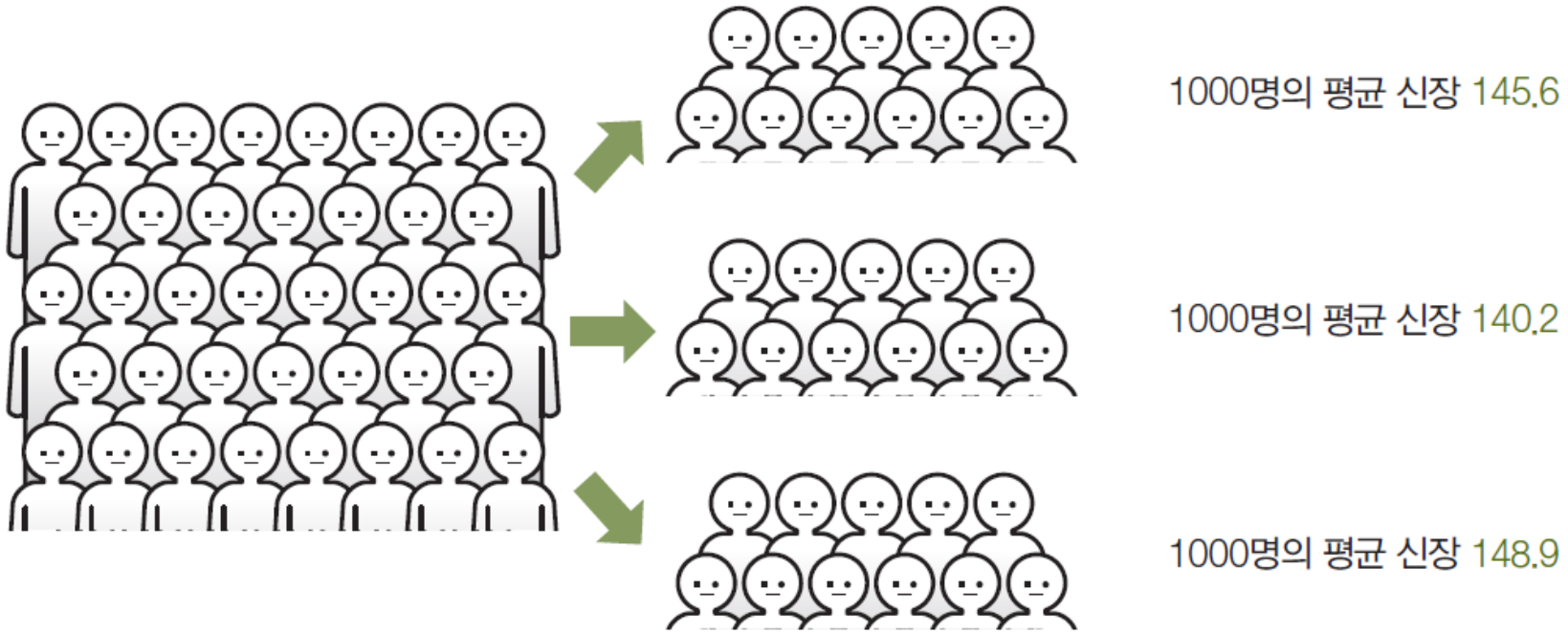
$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

  - 통계량 (statistic) : 표본의 특성을 나타내는 수치
    - 표본 분포 (sampling distribution) : 표본으로 부터 얻은 통계량의 확률 분포
      - 표본평균, 표본분산, 표본표준편차, 표본비율



추론 통계 – 모집단과 표본

- 표본
  - 선택된 일부는 우연히 선택된 것으로, 선택될 때마다 내용이 바뀜
  - 추측통계학의 전제 조건은 임의(무작위) 표본 선택

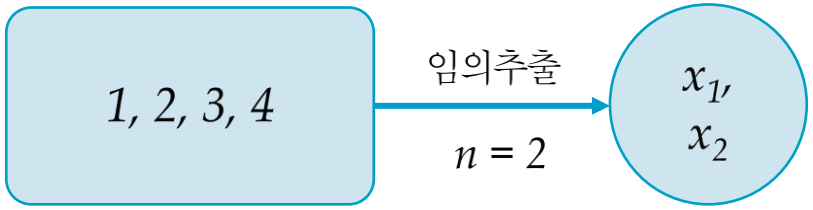


추론 통계 – 모집단과 표본

- 예제
  - 모평균 $\mu = 1\times\frac{1}{4}+2\times\frac{1}{4}+3\times\frac{1}{4}+4\times\frac{1}{4}=\frac{5}{2}$
  - 모분산 $\sigma^2 = \frac{1}{4}(1^2+2^2+3^2+4^2)-\left(\frac{5}{2}\right)^2=\frac{5}{4}$

X	1	2	3	4
$P(X=x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

- n=2인 표본평균의 확률 분포



가능한 표본	$\overline{X}$	가능한 표본	$\overline{X}$
$\{1,1\}$	1	$\{3,1\}$	2
$\{1,2\}$	1.5	$\{3,2\}$	2.5
$\{1,3\}$	2	$\{3,3\}$	3
$\{1,4\}$	2.5	$\{3,4\}$	3.5
$\{2,1\}$	1.5	$\{4,1\}$	2.5
$\{2,2\}$	2	$\{4,2\}$	3
$\{2,3\}$	2.5	$\{4,3\}$	3.5
$\{2,4\}$	3	$\{4,4\}$	4

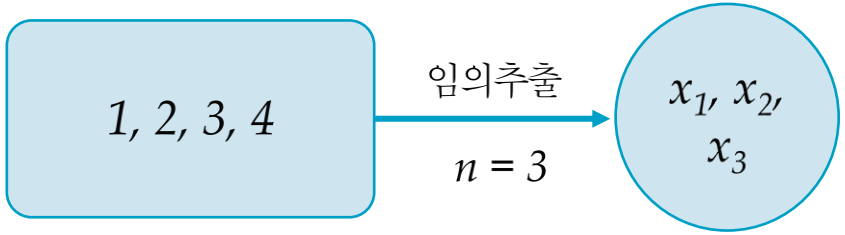
추론 통계 – 모집단과 표본

- 예제

$\overline{X}$	1	1.5	2	2.5	3	3.5	4	합계
$P(\overline{X}=x)$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	1

  - 표본평균 $\mathrm{E}(\overline{X}) = \mu = \frac{5}{2}$
  - 표본분산 $\mathrm{Var}(\overline{X}) = \frac{5/4}{2} = \frac{5}{8}$

- n=3인 표본평균의 확률 분포



$\overline{X}$	1	$\frac{4}{3}$	$\frac{5}{3}$	2	$\frac{7}{3}$	$\frac{8}{3}$	3	$\frac{10}{3}$	$\frac{11}{3}$
$P(\overline{X}=\overline{x})$	$\frac{1}{64}$	$\frac{3}{64}$	$\frac{6}{64}$	$\frac{10}{64}$	$\frac{12}{64}$	$\frac{12}{64}$	$\frac{10}{64}$	$\frac{6}{64}$	$\frac{3}{64}$

  - 표본평균 $\mathrm{E}(\overline{X}) = \mu = \frac{5}{2}$
  - 표본분산 $\mathrm{Var}(\overline{X}) = \frac{5/4}{3} = \frac{5}{12}$

$$\mu_{\overline{X}} = \mu, \quad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

추론 통계 – 표본 추출

유명 커피전문점 테이크아웃 커피 비교

브랜드	사이즈	아메리카노		캐러멜 마키아토	
		매장별 최대 용량 편차(g)	카페인 함유량 (㎎/1잔)	매장별 최대 용량 편차(g)	카페인 함유량 (㎎/1잔)
스타벅스	톨	60	114	107	66
커피빈	스몰	77	168	51	83
파스쿠찌	레귤러	52	196	81	116
카페베네	레귤러	46	168	77	84
엔제리너스	스몰	56	95	78	90
탐앤탐스	톨	61	91	60	104
투썸플레이스	레귤러	83	159	113	80
할리스커피	레귤러	55	152	131	145
이디야커피	원 사이즈	54	91	79	90

추론 통계 – 표본 추출

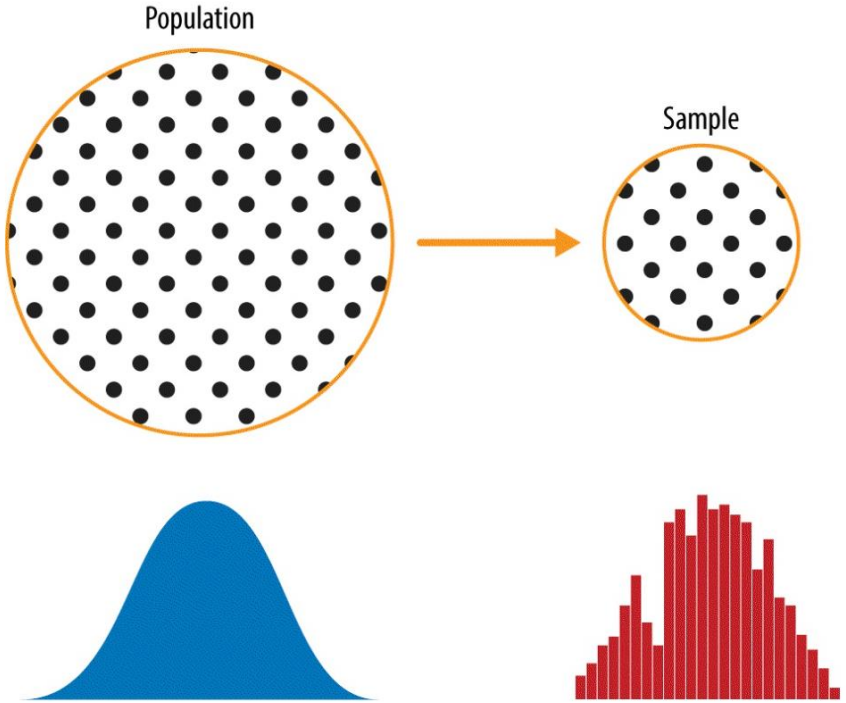
- 모집단과 표본의 중요성
  - 조사 불가능 : 모집단 전체를 대상으로 조사함은 불가능
  - 시간과 비용 : 모집단을 다 조사하는 데는 많은 시간과 비용 소요
  - 표본 추출 (sampling)을 통해 이를 해결 → 표본 추출은 데이터 분석의 난제를 일부 해결 가능

- 주의할 점
  - 우리가 실제로 알고 싶은 값은 표본의 값 (통계량 statistic)이 아니고 모집단의 값 (모수 parameter) 임
  - 추론 통계학의 목적 : 추론 (inference) → 표본에서 구한 값을 이용해 우리가 구하고자 하는 모수도 이럴 것이다 라고 유추
  - 편의 (bias): 데이터로부터 발견한 어떠한 추론도 해당 사용자 집단 또는 어떤 특정한 날의 사용자들 이외에 다른 사람들에 대한 일반화된 결론으로 확대되는 것을 경계해야 함

- 통계학에서는 모집단과 표본의 관계를 수학적으로 모형화

추론 통계 – 일반화

- 일반화 (Generalization)
  - 추출된 표본이 올바른 표본 추출방법으로 추출되었기 때문에 이 표본을 통해서 구한 값이 모집단의 값을 대표할 수 있다
  - 표본의 대표성 (representativeness)
  - 보통 분석 보고서의 앞부분에 기술 통계를 이용해 표본의 특성 및 연구 조사방법에 대해 서술
  - 이 부분을 통해 우리는 이 분석이 올바로 진행이 되었는지를 알 수 있고, 해당 결과를 일반화할 수 있는 지 판단할 수 있음



추론 통계 – 표본 오차

- 표본 오차 (sampling error)
  - 조사
    - 전수조사 (complete survey) : 모집단 전체를 대상으로 조사
    - 표본조사 (sampling survey) : 모집단의 일부인 표본만 대상으로 조사

  - 모집단에서 표본을 추출해서 조사하기 때문에 모수와 통계량 사이에 생기는 오차
  - 표본 오차는 아무리 표본을 크게 해도 전수조사를 하지 않는 이상 존재
  - 표본의 크기를 크게 함으로써 표본 오차를 최소화
  - 가설 검정 : 표본 오차의 허용 범위를 확률로 구하는 것

추론 통계 – 표본 조사 방법

- 확률 추출
  - 모집단에 속하는 모든 추출 단위에 대해 사전에 일정한 추출 확률이 주어지는 표본 추출법
    - 단순 임의 추출법 (Simple Random Sampling) :N개 중 n개의 표본을 추출할 때 모든 같은 확률로 추출
    - 계통 추출법 (systematic sampling) : 추출할 데이터를 차례로 나열 후, 일정 간격을 두고 추출
    - 층화 추출법 (stratified sampling) : 모집단을 몇 개의 동질적인 층으로 나눈 후, 각 층에서 SRS
    - 군집 추출법 (cluster sampling) : 모집단을 소집단들로 나누고, 일정 수의 소집단을 무작위로 표본추출 후, 추출된 소집단 내에서 전수조사

- 비확률 추출
  - 추출단위가 표본에 추출될 확률을 객관적으로 나타낼 수 없는 표본 추출법
  - 인위적으로 표본 추출하며, 변수 수집 비용이 적게 들지만 분석 결과 일반화 불가능
    - 편의 표본 추출 (convenience sampling) : 조사자 편의에 따라 모집단으로부터 편리한 방법을 통해 추출
    - 판단 표본 추출 (purpose sampling) : 조사자의 주관에 따라 표본의 대상을 선정
    - 할당 표본 추출 (quota sampling) : 층화 표본 추출법과 같이 소집단을 선택한 후 그 안에서 작위적으로 추출

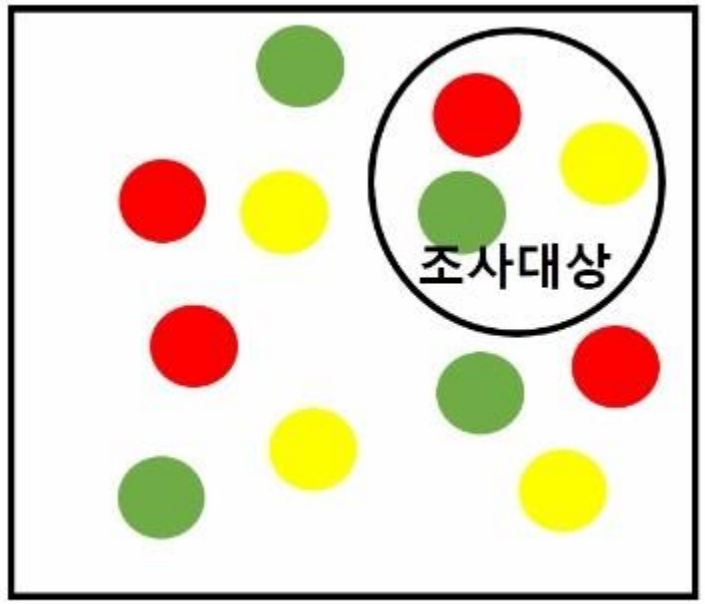
추론 통계 – 표본 조사 방법

- 확률표본추출과 비확률표본추출 비교

	확률표본추출	비확률표본추출
유형	단순무작위표본추출, 층화표본추출, 군집표본추출	편의표본추출, 판단표본추출, 할당표본추출
표본추출 방법	무작위적 표본추출	인위적 표본추출
모수추정의 편의(bias) 여부	모수추정에 편의 없음	모수추정에 편의 존재
표본분석의 일반화 여부	일반화 가능함	일반화를 하는데 제약 있음
표본오차의 추정 여부	추정 가능함	추정 불가능함
표본추출하는 데 소요되는 시간 및 비용	많이 소요됨	적게 소요됨
사용하는 경우	연구대상이 표본으로 추출될 확률이 알려져 있을 때	연구대상이 표본으로 추출될 확률이 알려져 있지 않을 때

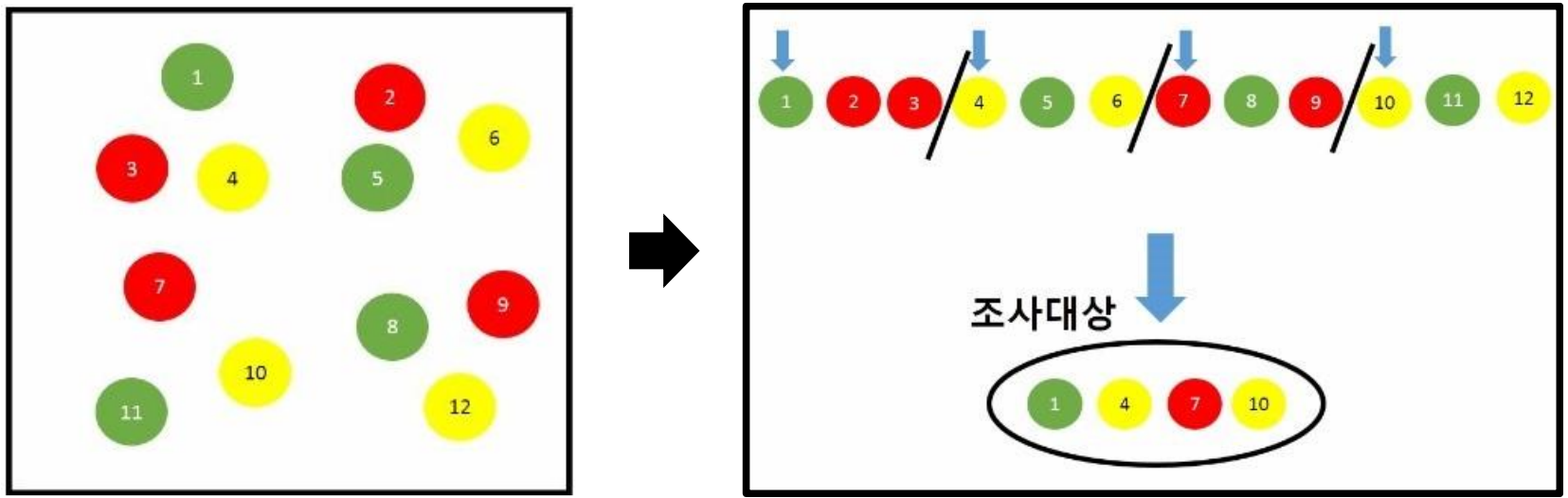
추론 통계 – 표본 조사 방법

- 확률 추출
  - 단순 임의 추출법 (Simple Random Sampling)
    - 모집단 전체의 일련번호를 부여해서 표본조사 틀을 만든 후, 난수표 등을 이용하여 각 객체가 뽑힐 가능성이 동일하게 되게끔 표본을 추출하는 방법
    - 예시 : 제비 뽑기, 사다리 타기, 20면체 주사위 던지기, …
    - 장점
      - 모집단에 대한 사전 지식 불필요
      - 추출 기회가 동등하고 독립적이기 때문에 표본의 대표성이 높음
      - 비용절감 및 시간단축
    - 단점
      - 모집단에 대해 갖고 있는 지식을 활용할 수 없음
      - 비교적 표본의 규모가 커야 함



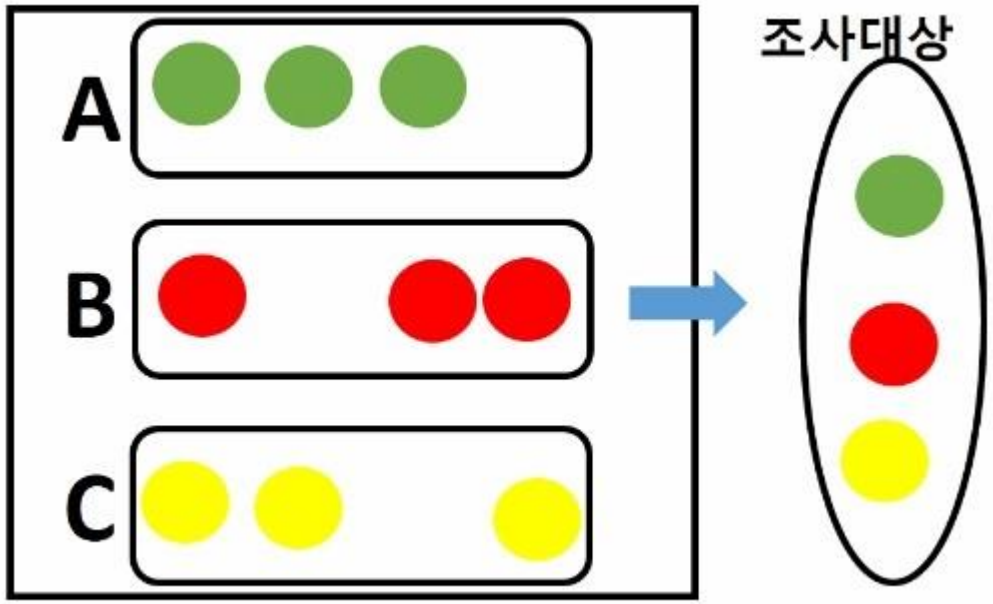
추론 통계 – 표본 조사 방법

- 확률 추출
  - 계통 추출법 (systematic sampling)
    - 모집단을 그룹으로 나누고 각 그룹에서 n번째 간격마다 하나씩의 단위를 표본으로 추출
    - 예시 : 경품을 10, 20, 30, 번째 고객에게 증정
    - 장점
      - 표본 추출이 간편하며 단순임의추출법의 대용으로 자주 사용
      - 일반적으로 표본이 모집단 전체를 잘 반영
    - 단점
      - 표본의 대표성이 저해 (주기성, 특정 경향성을 보인다면 피할 것)



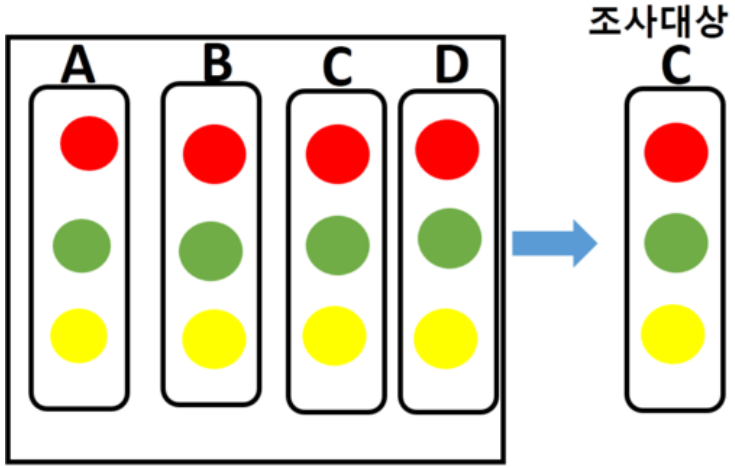
추론 통계 – 표본 조사 방법

- 확률 추출
  - 층화 추출법 (stratified sampling)
    - 모집단을 먼저 서로 겹치지 않는 여러 개의 그룹으로 분할 후, 각 층에서 단순임의추출법에 따라 추출
    - 모집단을 어떠한 기준에 따라 분류해야 함
    - 장점
      - 표본의 크기가 크지 않아도 대표성 보장
      - 각 층화 집단의 특수성을 알 수 있고 비교 가능
    - 단점
      - 모집단의 각 층에 대한 정보 필요
      - 층화가 너무 복잡한 경우 오히려 표본오차 증가
      - 시간과 비용 증가



추론 통계 – 표본 조사 방법

- 확률 추출
  - 군집 추출법 (cluster sampling)
    - 모집단을 다수의 이질적인 성격을 가진 다수의 집락 (군집)으로 구분하고 , 추출된 군집 내의 일부 또는 전체를 조사하는 방법
    - 예시 : 서울에 있는 대학생 월 평균 용돈 추정
      - 모집단 : 서울에 있는 약 60개의 대학교 학생 전체
      - 서울에 있는 대학 리스트 이용해 무작위로 몇 개 대학 추출 후 계열, 학과, 학년에 맞게 대학생 추출
    - 장점
      - 모집단의 분포가 널리 퍼저있는 전국 단위 대규모 조사에 이용
      - 모집단의 목록이 없어도 가능
    - 단점
      - 집단 내부 이질성 확보가 어려움
      - 다른 방법 대비 오차가 더 커질 수 있음



추론 통계 – 표본 조사 방법

- 비확률 추출
  - 판단 표본 추출 (purpose sampling)
    - 조사 문제를 잘 알고 있거나 모집단의 의견을 반영할 수 있을 것으로 판단되는 특정 집단을 추출
    - 전문적 지식을 가진 표본을 찾아내는 것이 매우 중요
    - 예시 : 동덕여대 앞 커피 전문점 개점을 위한 사전 조사 시, 주변 점포 주인들 대상 조사
    - 장점
      - 적은 비용으로 의미 있는 변수 수집 가능
      - 선정된 표본이 실제로 대표성을 가지는 경우 매우 효과적
    - 단점
      - 모집단에 대한 충분한 지식 필요
      - 연구자가 특정의 요소가 대표성이 있다고 믿는다 하여 실제 그렇다는 보장이 없음

추론 통계 – 표본 조사 방법

- 비확률 추출
  - 할당 표본 추출 (quota sampling)
    - 미리 정해진 분류기준에 의해 모집단을 여러 소집단으로 구분하고 각 집단별로 필요한 대상 추출
    - 모집단의 특성을 잘 반영할 수 있는 표본집단을 구성할 수 있어서 일반적으로 가장 널리 사용
    - 일반적인 설문 면접 대상자 선정 : 연령, 성별, 소득 등에 의해 표본 할당
    - 장점
      - 비확률 표본추출 방법 중 가장 정교함
      - 높은 수준의 대표성 확보 (시간, 경제적인면 장점)
    - 단점
      - 조사자의 주관적인 판단 개입 (한계)

가설 검정

- (통계적) 가설 : 모집단의 미지의 모수에 대한 주장이나 서술
  - 대립가설 (alternative hypothesis, H1)
    - 입증하여 주장하고자 하는 가설
  - 귀무가설 (null hypothesis, H0)
    - 대립가설의 반대 가설
    - 대립가설을 입증할 수 없을 때 무효화 시키면서 받아들이는 가설
  - 예시
    - 어느 과자의 겉봉지에 용량이 150g이라 표시되어 있는데 실제 열어보니 이보다 적음
      - 대립가설 (H1) : 과자의 용량이 150g 미달한다.
      - 귀무가설 (H0) : 과자의 용량이 150g 이다.
    - 휴대폰의 색상에 따라 여성과 남성의 선호도가 같은 지 알아봄
      - 대립가설 (H1) : 휴대폰의 색상에 따라 남녀의 선호도는 다르다.
      - 귀무가설 (H0) : 휴대폰의 색상에 따라 남녀의 선호도는 같다.

가설 검정

- 가설 검정의 결론과 오류의 종류

사실 \ 결과	귀무 가설 채택	귀무 가설 기각
귀무 가설이 맞음	옳은 결론	잘못된 결론 (제 1종 오류)
귀무 가설이 틀림	잘못된 결론 (제 2종 오류)	옳은 결론

  - 제 1종 오류 (α error) : 귀무가설 H0가 옳은데도 불구하고 H0를 기각하게 되는 오류
    - 제 1종 오류를 범할 확률의 최대 허용한계를 유의 수준 (significance level) 또는 위험율이라고 함
    - 현재 표본으로부터 얻어진 정보로는 대립가설이 사실임을 입증할 만한 충분한 증거가 없음을 의미
  - 제 2종 오류 (β error) : 귀무가설 H0가 옳지 않은데도 H0를 채택하게 되는 오류

  - 일반적으로 유의수준 α를 미리 정해 놓고, β를 최소화하는 검정의 방법을 사용

가설 검정

- 예시 : 가위바위보
  - 김태완이라는 사람이 자기는 가위/바위/보 게임을 잘 한다고 주장
  - 그 주장을 입증하기 위해서 실험을 하였고, 두 번 해서 두 번 다 이겼다. 그럼 정말 나는 가위/바위/보를 잘 하는 사람이라고 결론 지을 수 있을까?
  - 만약 열 번해서 열 번 다 이겼다면 그 결론은 바뀌나?

  - 귀무 가설 : 가위바위보 게임의 승률 (p=0.5)
  - 대립 가설 : p > 0.5
  - 우연히 두 번 다 이길 확률은 0.5*0.5 = 0.25 → 귀무 가설 기각 X
  - 우연히 열 번 다 이길 확률은 (0.5)^10 = 1/1024 (0.1%) → 귀무 가설 기각, 그리고 대립 가설 채택

    귀무 가설 하에서 나타나게 될 데이터의 확률 분포 상에서 이 사건이 얼마나 극단적인 것인가를 판단하는 척도 (p-value)

  - 주로 p 값은 5%가 기준이 되나 종종 1%나 10%도 있음 → 유의 수준 (significance level)
    - p-value가 유의 수준보다 작으면 관측 값이 유의 (significant)한 것이 되고, 이 경우 귀무 가설 기각

가설 검정

- 오류의 가능성은 존재
  - 김태완이 정말 우연히 열 번을 다 이겼을 수도 있음 → 귀무 가설 기각 → 제 1종 오류
    - 귀무 가설이 참임에도 불구하고 그것을 기각
  - 제 1종 오류를 낮추기 위해서는 유의 수준을 낮추면 됨 → 제 2종 오류가 커짐
  - 제 1종 오류를 작게 하는 것이 좋음
  - 만약 표본수를 늘리면 두 오류를 동시에 줄일 수 있음

회귀 분석 – 역사

- Historical Origin of the Regression (회귀)

- 회귀 (regression)라는 용어는 유전학자 Francis Galton (1886)에 의해 처음 사용

- 그의 논문에서 “비정상적으로 크거나 작은 부모의 아이들 키는 전체 인구의 평균 신장을 향해 움직이거나 회귀하는 경향이 있다.” 주장

- 그의 친구 Karl Pearson (1903)은 1,000명 이상의 변수를 수집하여 Galton의 보편적 회귀의 법칙 (law of universal regression)을 다음과 같이 확인

- 키가 큰 아버지 집단의 아들 평균 신장은 아버지 보다 키가 작았고, 키가 작은 아버지 집단의 아들 평균 키는 아버지보다 컸다. 즉, 아들의 키는 아버지의 키와 상관없이 전체 남자들의 평균 신장을 향해 회귀한다는 것임

회귀 분석 – 기본 용어

- 상수 (constant)와 변수 (variable)
  - 상수 : 항상 같은 수
  - 변수 : 변하는 수

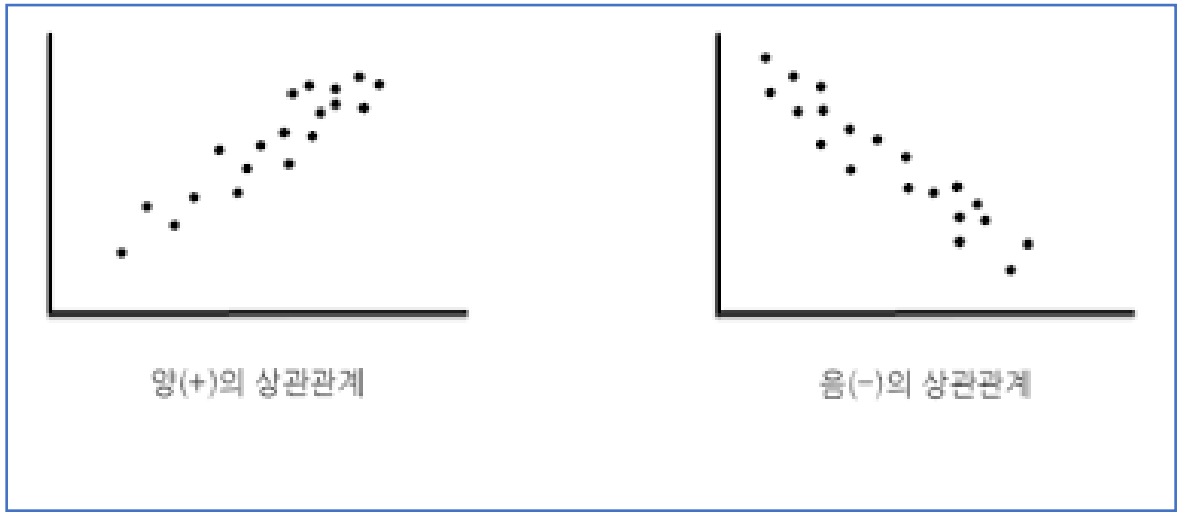
- 독립변수
  - 예측변수, 원인변수, 설명변수, 입력변수
- 종속변수
  - 준거변수, 결과변수, 반응변수, 출력변수

- 연속형 변수와 범주형 변수

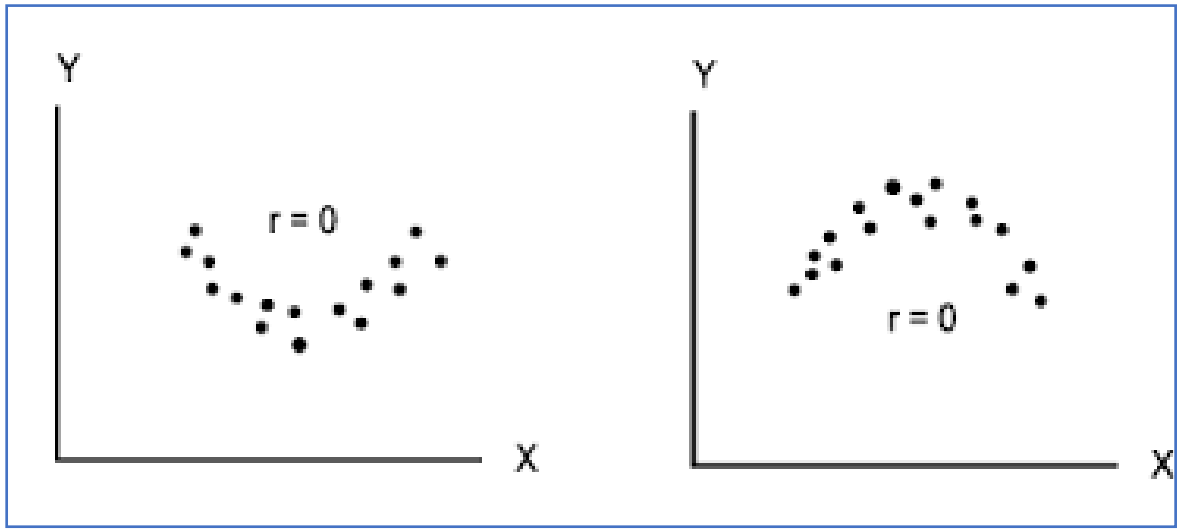
변수 간의 관계

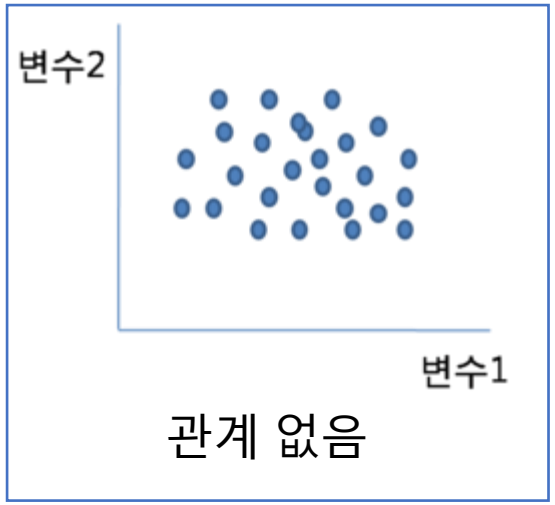
- 한 변수가 변할 때 다른 변수는 어떻게 변화하는가?
  - 예시 : 학습 동기와 성취도, 불안과 성취도
  - 선형 관계 :

Scatter plot (산포도)



  - 비선형 관계 :





상관 관계

- 한 변수와 다른 변수가 공변하는 함수관계
  - 양의 상관관계 : 두 변수가 같은 방향으로 움직임
  - 음의 상관관계 : 두 변수가 다른 방향으로 움직임

- 상관 계수
  - 상관계수는 계산식은 복잡하나 -1 ~ 1 사이의 값을 가짐
  - 상관계수가 -1일 때 완벽한 음의 상관관계
  - 상관계수가 0일 때 아무런 관계가 없음
  - 상관계수가 1일때 완벽한 양의 상관관계
  - 상관계수의 +/- 는 방향을 의미
  - 상관계수의 크기는 힘을 의미
    - 상관계수가 절대값 1에 가까울 수록 힘이 세다는 의미
    - 힘이 세다는 것은 데이터들이 가깝게 모여 있다는 의미
    - 따라서 데이터들이 퍼져 있으면 상관계수가 0에 가까워 짐

감사합니다

kimtwan21@dongduk.ac.kr

김 태 완