3. 확률과 분포

추론통계와 확률

- 추론: 알고 있는 정보를 근거로 다른 판단을 도출하는 과정
- 통계적 추론:
  - 부분의 통계 값을 이용하여 전체의 특징을 추정하는 행위
  - 기술 통계와 달리, 일어나지 않은 상황을 예측해야 함
- 확률
  - 사건이 일어날 가능성 계산
  - 부분을 통해 전체를 추측
  - 추측 결과의 신뢰 가능성 판단
  - 추정치를 바탕으로 그 다음 행동 계획 수립

확률

- 확률(Probability)
  - 실험이나 관측 가능한 상황에서 특정 사건의 발생 가능성
  - 특정 사건이 발생할 경우의 수를 모든 가능한 경우의 수로 나눈 값
  - 불확실성을 정량화하는 데 필요한 도구
- 확률 – 통계
  - 확률: 개별 사건의 발생 경향성
  - 통계: 여러 사건들이 발생하는 경향성, 인과성, 상관성 등을 파악
  - Ex) 전염병

https://newsimg-hams.hankookilbo.com/

확률 계산 방법

- 고전주의 확률(빈도주의, Frequentist P)
  - 빈도를 셈
  - 장기간 관측을 통하여 정확한 값 추정
  - Why? 세상이 조건이 동일한 시행은 없음
  - 동전을 100 번 던졌을 때 앞면이 나올 확률
  - 하지만
    - 데이터가 없거나 부족하다면?
    - 발생 가능한 모든 경우를 상정하기 어렵거나, 그 값이 불확실하다면?
    - 서로 다른 실험 간 결과가 일치하지 않는다면?

https://www.theglobeandmail.com/globe-investor/inside-the-market/how-a-simple-coin-toss-game-can-make-you-a-better-investor/article33183678/

확률 계산 방법

- 베이지언 확률(Bayesian P)
  - 확률 = 믿음
    - 특정 사건이 특정 순간에 발생하는 경우에 대한 믿음의 정도
    - 객관성에 대한 관측자들 간 합의가 필요
  - 사건 발생에 관련된 지식이 추가될 경우 확률은 바뀔 수 있음
    - 슈퍼컴퓨터의 월드컵 우승 예측
  - 주관주의, 도박

https://www.theglobeandmail.com/globe-investor/inside-the-market/how-a-simple-coin-toss-game-can-make-you-a-better-investor/article33183678/

확률 계산 방법

- 샘플 공간(Sample space)
  - 모든 발생 가능한 결과들의 집합
  - 전사건(S)
- 사건(Event)
  - 샘플 공간의 부분 집합
  - 특정 조건을 만족하는 시행(Experiment)들의 모임
- 확률
  - $P(A) = \frac{n(A)}{n(S)}$
  - 복권에 당첨될 확률 $= \frac{\text{추첨된 수의 조합}}{\text{복권에서 고를 수 있는 모든 경우}}$

1등 계산

$$\frac{(6C6)*(39C0)}{(45C6)} = \frac{1*1}{8,145,060} = \frac{1}{8,145,060} = 0.00001227\%$$

2등 계산

$$\frac{(6C5)*(38C0)*(1C1)}{(45C6)} = \frac{6*1*1}{8,145,060} = \frac{6}{8,145,060} = 0.00007366\%$$

3등 계산

$$\frac{(6C5)*(38C1)*(1C0)}{(45C6)} = \frac{6*38*1}{8,145,060} = \frac{228}{8,145,060} = 0.00279924\%$$

4등 계산

$$\frac{(6C4)*(39C2)}{(45C6)} = \frac{15*741}{8,145,060} = \frac{11,115}{8,145,060} = 0.13646308\%$$

5등 계산

$$\frac{(6C3)*(39C3)}{(45C6)} = \frac{20*9,139}{8,145,060} = \frac{182,780}{8,145,060} = 2.24405958\%$$

https://m.blog.naver.com/wintergalaxy/221450099192

확률 계산 방법

- 독립 사건(Independent event)
  - 두 사건이 서로 영향을 미치지 않는 경우
  - 사건 A의 발생이 사건 B의 발생 확률에 영향을 주지 않음
  - 즉석복권 vs 복권
  - $P(A \cap B) = P(A) \times P(B)$
- 종속 사건(Dependent event)
  - 한 사건이 다른 사건의 발생 확률에 영향을 미치는 경우

https://www.hani.co.kr/arti/sports/

빈도주의 – 베이지언

- Ex)주사위를 두 번 던질 때 두 수가 같을 확률?
  - 샘플 공간: 주사위를 두 번 던져 나올 수 있는 모든 수의 조합 = {(1, 1), … , (6, 6)}
  - 사건: 주사위를 두 번 던져 나온 수(2, 5)
- 확률 계산법
  - 빈도주의: N번 던져 결과를 확인
  - 베이지언:
    - 실험 전: 대상의 발생 가능성을 추론하여 추정
      - 사전 확률(Prior)
    - 실험 후: 실험 중 발생한 기타 요소들을 모두 고려하여 확률 수정
      - 사후 확률(Posterior)

큰 수의 법칙

- 빈도주의의 단점
  - 실험의 횟수가 제한
  - 예측할 미래의 변동성이 지나치게 커짐
- 베이지언의 단점
  - 확률을 주관적으로 해석
  - 해석 방식이 합의되지 않을 경우 계산 신뢰 불가
- 큰 수의 법칙(Law of large numbers)
  - : 적당히 많은 횟수의 시행을 반복할 경우 정확하다고 가정

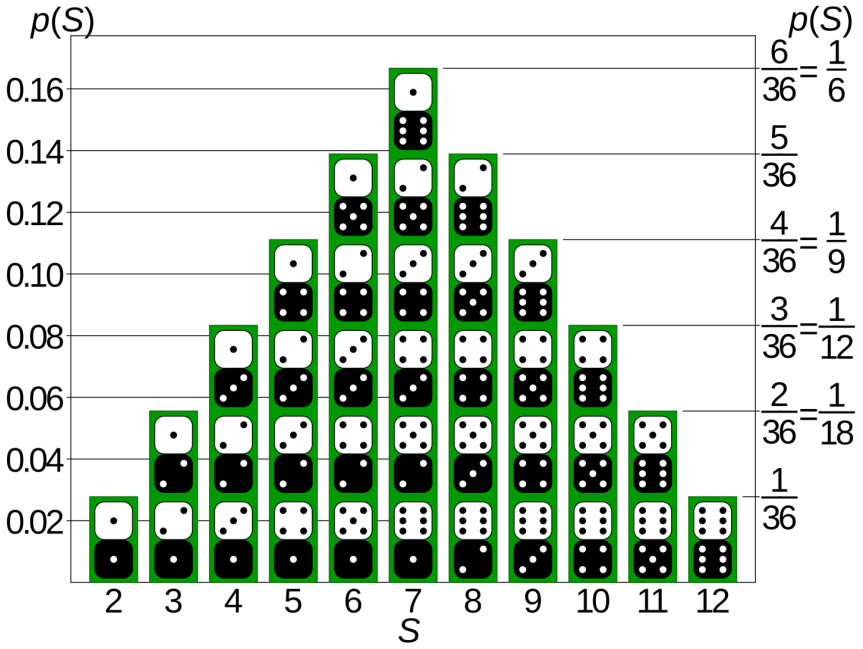
확률 변수

- 일어날 사건들과 그 확률을 함께 계산한 결과
  - 표본 공간에서 일어날 수 있는 사건들과, 그 확률을 연결
  - 불확실한 일을 고정하는 방법
  - Ex) 주사위를 두 번 던졌을 때 나온 두 눈의 합이 5일 경우 1000원을 얻는 게임
    - 주사위를 두 번 던져 나올 수 있는 수는 36가지
    - 두 눈의 합이 5인 경우 = (1, 4), (2, 3), (3, 2), (4, 1)
    - 두 눈의 합이 5일 확률 = 4 / 36 = 1 / 9
    - 단, 이 확률은 주사위를 무한 번 던졌을 때의 경우를 의미
  - 우리가 예상할 수 있는 수입
    - 1/9 * 1000 = 111.11원
- 기댓값: 특정 시행을 무한 번 반복했을 때 얻을 수 있는 값의 평균으로 기대할 수 있는 값
  - 확률의 평균

https://namu.wiki/w/%EA%B8%B0%EB%8C%93%EA%B0%92

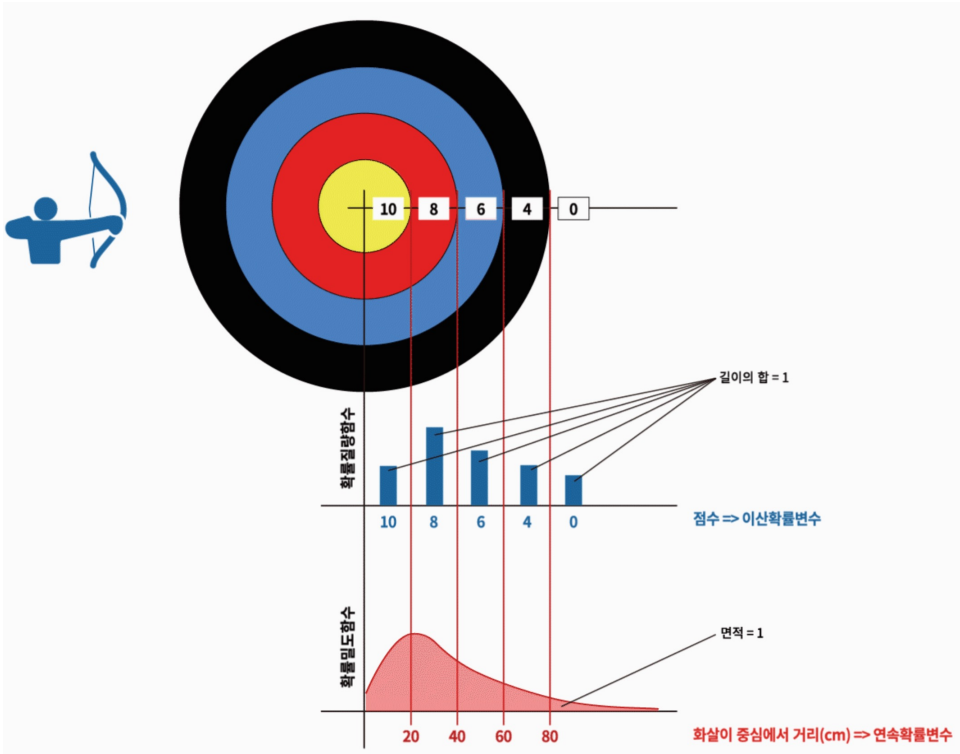
확률 분포

- 분포(Distribution)
  - 정해진 공간 안에서 자료들이 배치된 모습
  - 기술 통계 분포: 수집된 자료들이 놓여진 모습
- 확률 분포(Probability distribution)
  - 정해진 확률 공간 안에서 확률 변수들이 배치된 모습
- 이산 확률 분포(Discrete)
  - 이산형 확률 변수의 분포
  - 확률 질량 함수(P-Density Function)
    - 막대그래프



확률 분포

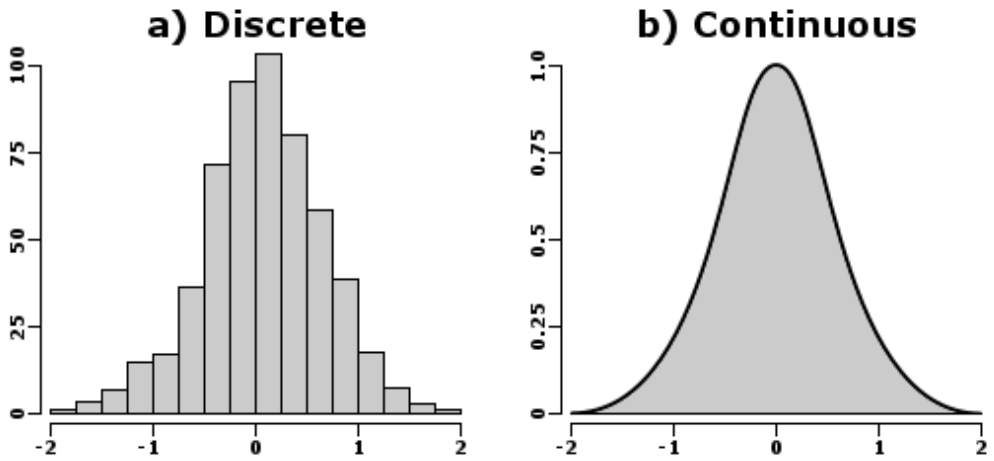
- 연속 확률 분포(Continuous PD)
  - 연속형 자료의 확률 분포
  - 데이터의 경계가 불분명하기에 이를 선으로 표현
  - 확률 밀도 함수(P-Mass Function)
- 특정 구간에서의 면적 = 해당 구간의 확률
  - 확률 질량 함수와 확률 밀도 함수 모두 해당

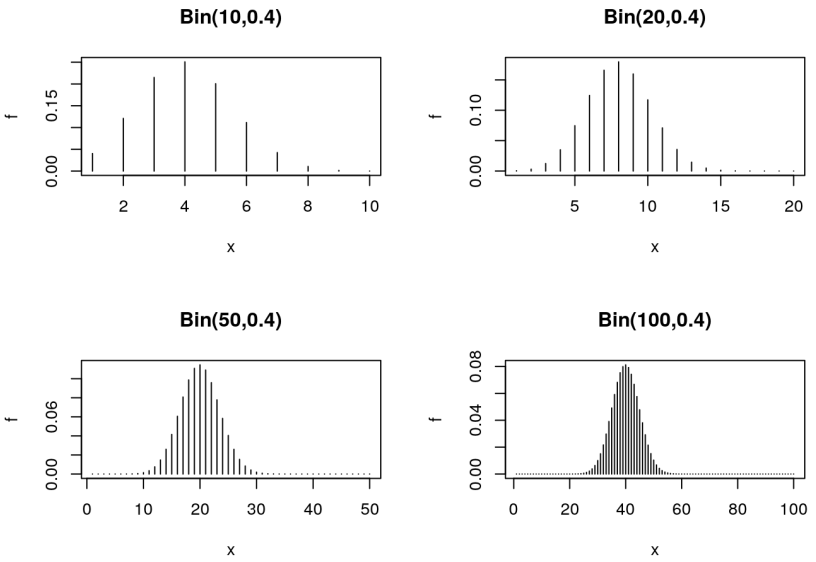


https://ai2it.tistory.com/125?category=1167949

확률 분포

- 이산 확률 분포와 연속 분포 변수의 관계
  - 이산 확률 분포의 구간이 아주 세밀하다면 연속 확률 분포에 가까워짐
  - 세상의 많은 연구 결과를 종합한다면 연속 확률 분포를 따르는 데이터가 압도적으로 많음

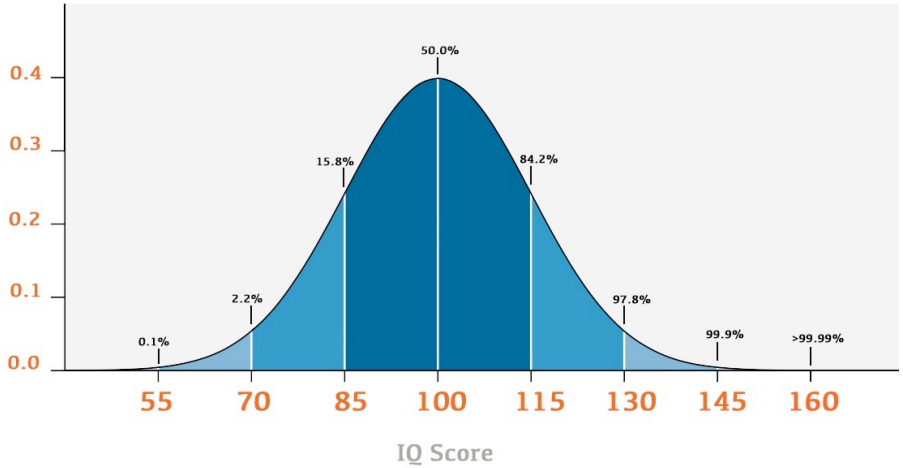




https://blog.firstpenguine.school/33, http://bigdata.dongguk.ac.kr/

중심극한정리

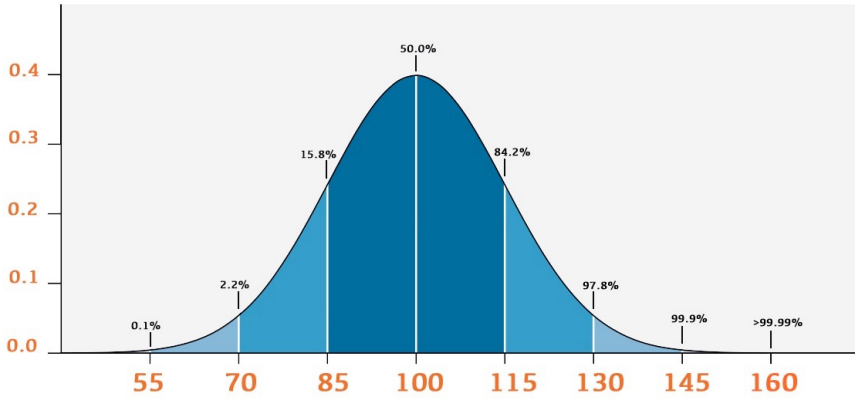
- 또한 수많은 통계 결과 종합 시
  - 데이터를 무작위로 추출했을 때 추출된 데이터의 크기가 충분히 크다면
  - 일반적인 사건이 많이 관찰되고, 극단적인 사건은 적게 관찰되며
  - 중심, 최빈값, 중앙값이 크게 차이나지 않은 분포(정규 분포, Normal distribution)를 표현함
  - 이러한 경향성을 중심극한정리(Central Linit Theorem)라 함

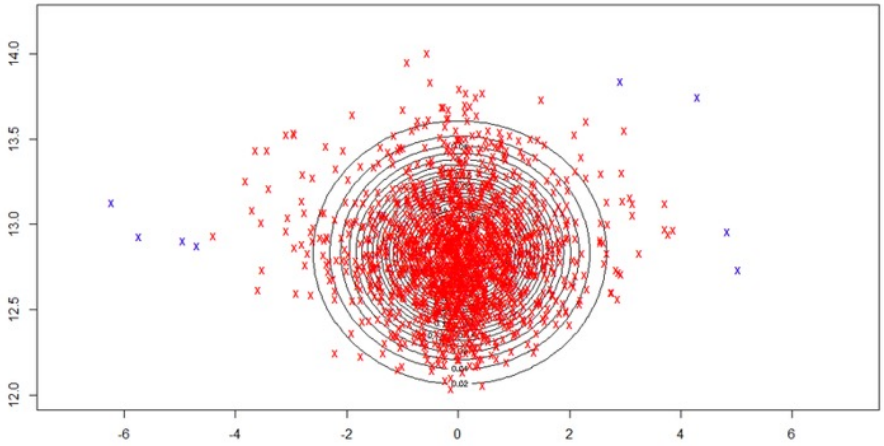


https://allthatvalue.com/column/67872

정규 분포

- 정규 분포의 특징
  - 가우스 분포(Gaussian distribution)라고도 함
  - 평균을 중심으로 한 종 모양의 대칭 그래프
  - 수많은 자연 현상과 사회 현상을 대변할 수 있는 일반적인 특징
  - 다루기 쉬움
    - 평균과 표준편차만으로 분포를 정확히 그릴 수 있음
    - 평균 = 중앙값 = 최빈값
  - 정규분포를 따르지 않는 데이터를 변환시킬 경우 비교 가능
    - 정규분포를 바탕으로 수많은 통계 추정 방법이 발달
    - 딥러닝도 그 중 하나



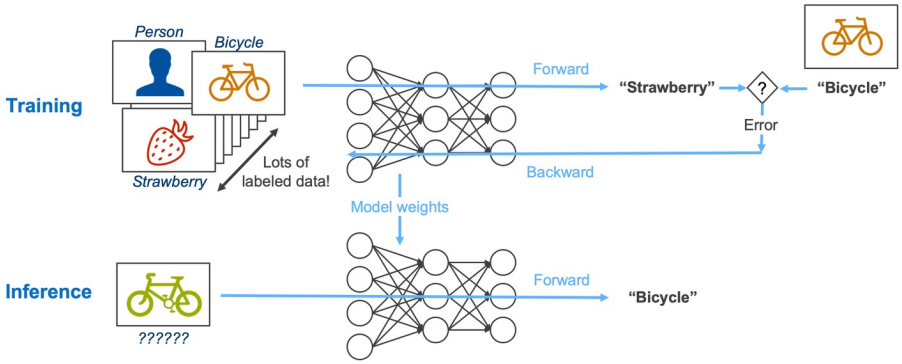


https://allthatvalue.com/column/67872

4. 추론 통계

추론 통계

- 추론 통계학(Inferential statistics)
  - 표본을 추출하여 모집단의 통계량을 추측하고
    - 모집단 추정
  - 오차를 감안하여 신뢰도를 관리하여
    - 구간 추정
  - 이를 바탕으로 특정한 가설을 세워 검증하는 학문
    - 가설 검정



https://m.dongascience.com/, https://www.linkedin.com/pulse/difference-between-deep-learning-training-inference-mark-robins-mdq8c/

표본과 모집단

- 모집단(Population)
  - 연구의 조사나 대상이 되는 전체 집합
  - 물리적인 전수조사 제한적

- 표본(Sample)
  - 모집단을 대표하는 작은 부분 집단
  - 모집단을 조사할 수 없어 일부만 관찰
  - 모집단의 대표성이 잘 드러나도록 추출해야 함

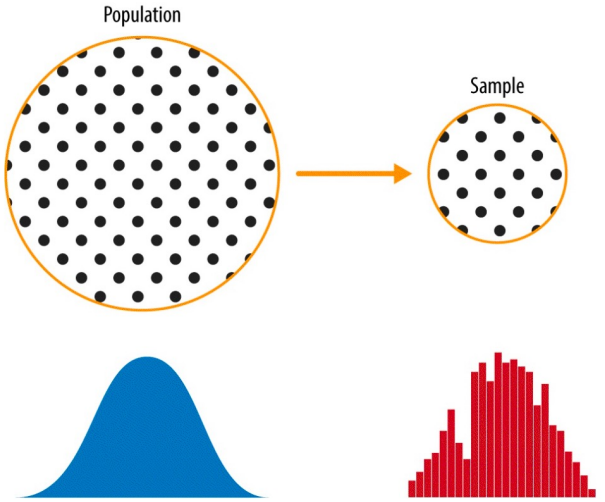


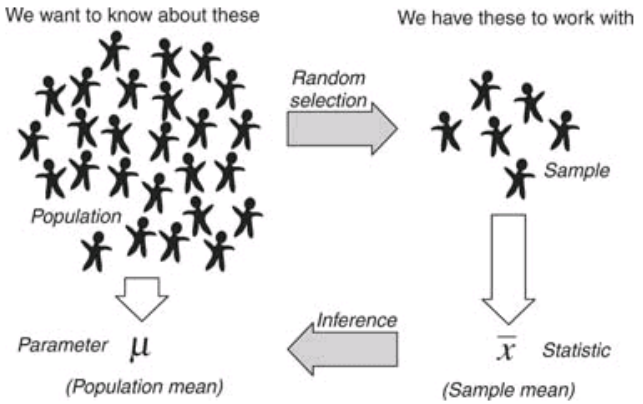
Figure 2-1. Population versus sample

모수와 표본의 통계량

- 모수(Parameter)
  - 연구자가 관심있는 모집단의 수치
  - 모평균, 모표준편차, 모분산, 모집단의 백분위수

- 통계량(Statistics)
  - 표본 자료에서 얻은 통계 수치
  - 표본 평균, 표분 분산, 표본 표준 편차 등

- 통계량을 이용하여 모수를 추정하는 것이 추론 통계학의 첫 번째 목표



http://bigdata.dongguk.ac.kr/lectures/med_stat/

표본 평균의 통계량

- 그러나 표본의 통계량은 모수와 어긋나는 경우가 많음
  - 추출된 표본이 편향될 가능성이 높음
  - 표본의 크기를 키우면 어느정도 해소
    - 그러나 이는 표본을 뽑는 취지에 반함

- 크기가 적당한 표본을 여러 번 뽑아 통계량을 낸다면?
  - 표본 평균의 통계량은 중심 극한 원리를 따름
  - 모집단이 정규분포를 따른다면?
  - 모평균 = 표본평균의 평균
  - 모표준편차 = 표본평균의 표준 편차

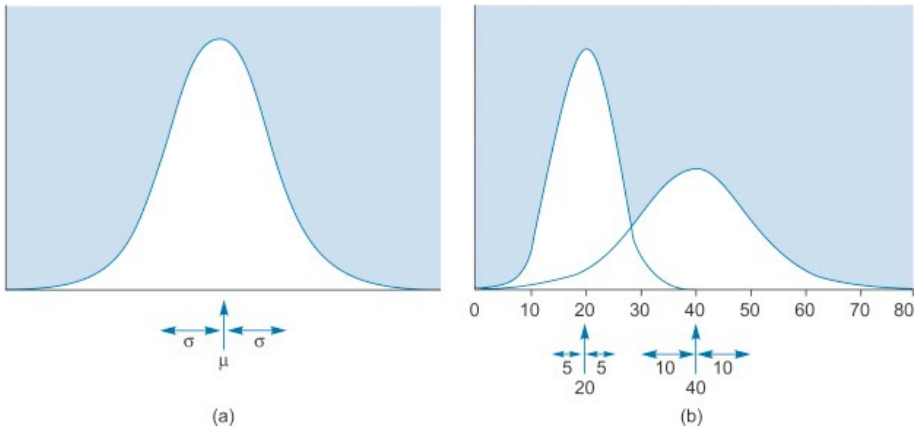
	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

https://ko.wikipedia.org/wiki/%EC%8B%AC%EC%8A%A8%EC%9D%98_%EC%97%AD%EC%84%A4

표준정규분포

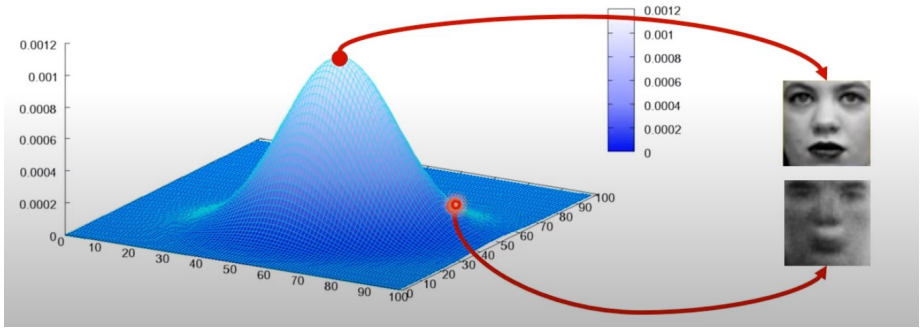
- 표준정규분포(Standard normal distribution)
  - 표준화된 정규분포
    - 평균이 0, 표준편차가 1인 정규분포
  - 평균과 표준편차가 서로 다른 두 정규분포를 비교하기 위한 방법
  - 정규분포를 따르지 않는 분포는 정규화와 표준화 모두 진행
  - 많은 통계적 추정 기법을 처리하기 위한 전처리 과정

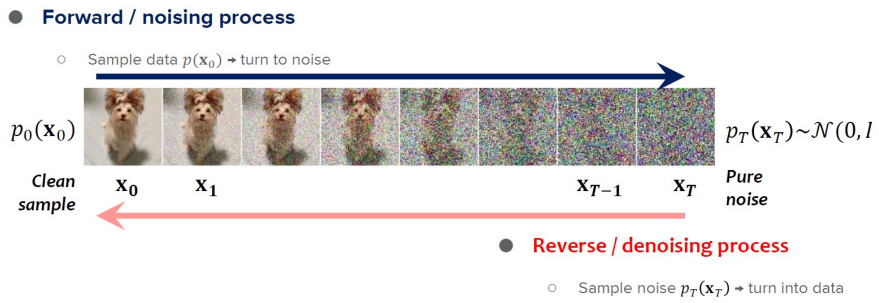


- 모집단이 정규분포를 따르지 않을 경우
  - 충분히 큰 크기로 여러 번 추출된 표본은 정규분포를 따름
  - 표본이 잘 뽑혔는지를 확인할 수 있음

정규분포와 머신러닝

- 머신러닝의 목표:
  - 학습 데이터를 바탕으로 일반화 성능을 높이는 것
  - 일반화 성능: 학습 데이터에 포함되지 않은 데이터도 인지할 수 있는 능력
  - 표본을 통해 모집단을 추정하는 추론 통계학의 연장선
- 수많은 EDA, 머신러닝 모델링과 학습 보조 알고리즘에 정규분포의 원리가 이용됨
  - 머신러닝 모델은 고급 통계분석기법에 기반
  - 데이터가 정규분포를 따를 수록 학습이 잘 됨





https://velog.io/@ym980118/, https://scholar.harvard.edu/binxuw/classes/machine-learning-scratch