AI 보안 기술개발 교육

머신러닝을 위한 통계

KISA 한국인터넷진흥원

AI 보안 기술개발 교육

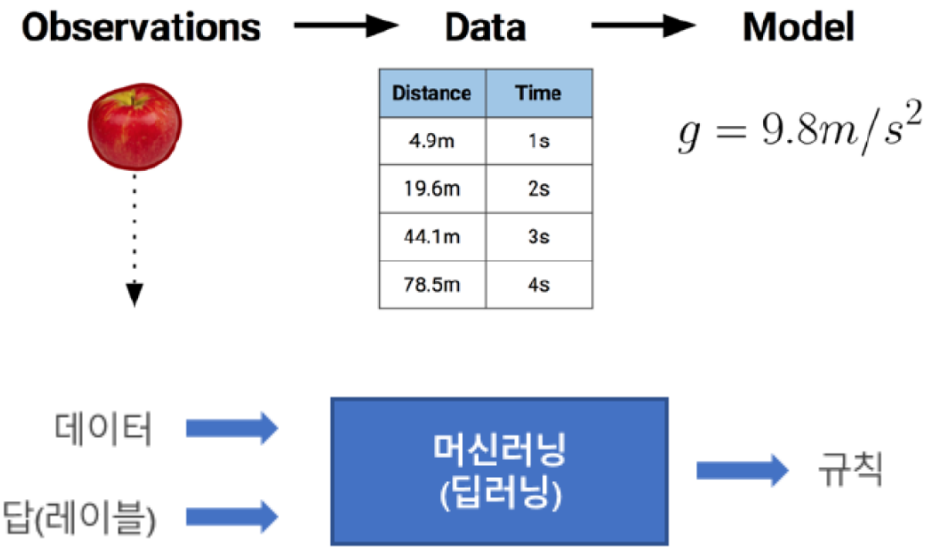
머신러닝을 위한 통계

1. 머신러닝 개요
2. 머신러닝의 분류
3. 통계 개요
4. 머신러닝을 위한 기초 통계

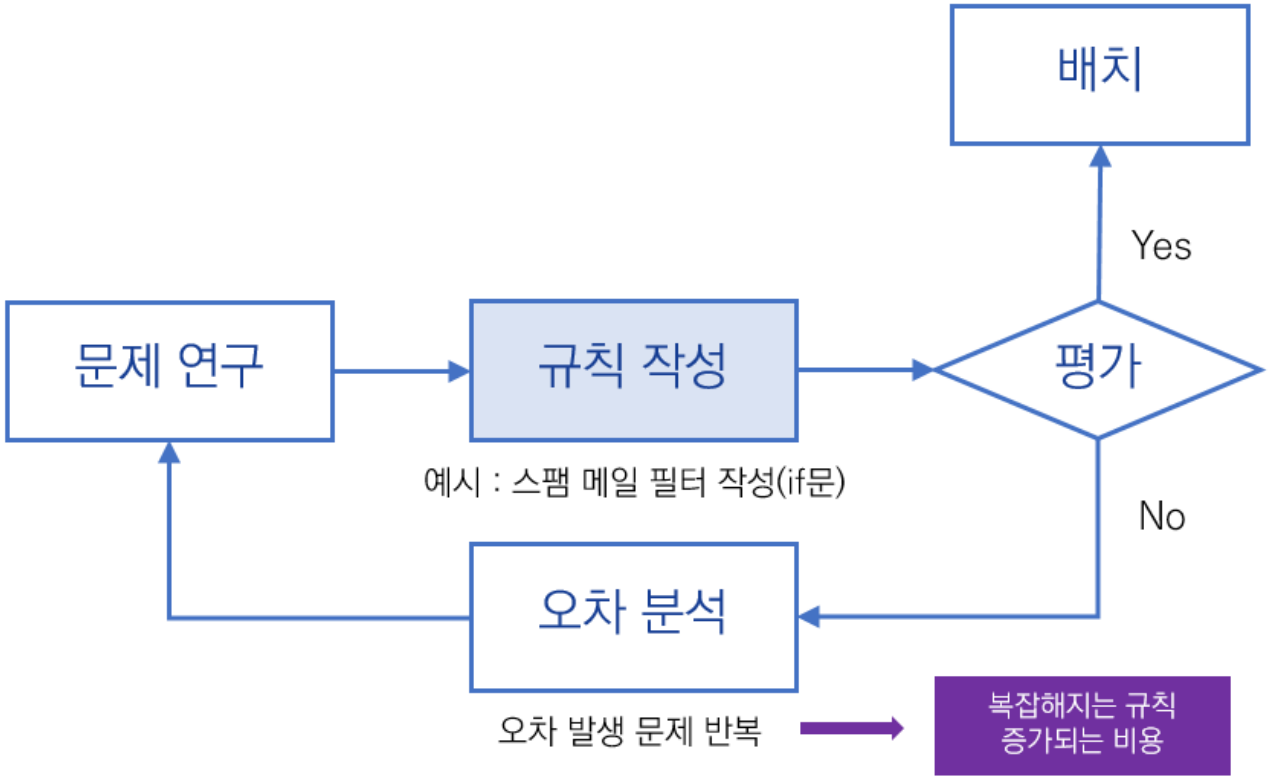
KISA 한국인터넷진흥원

머신러닝(Machine Learning) 개념

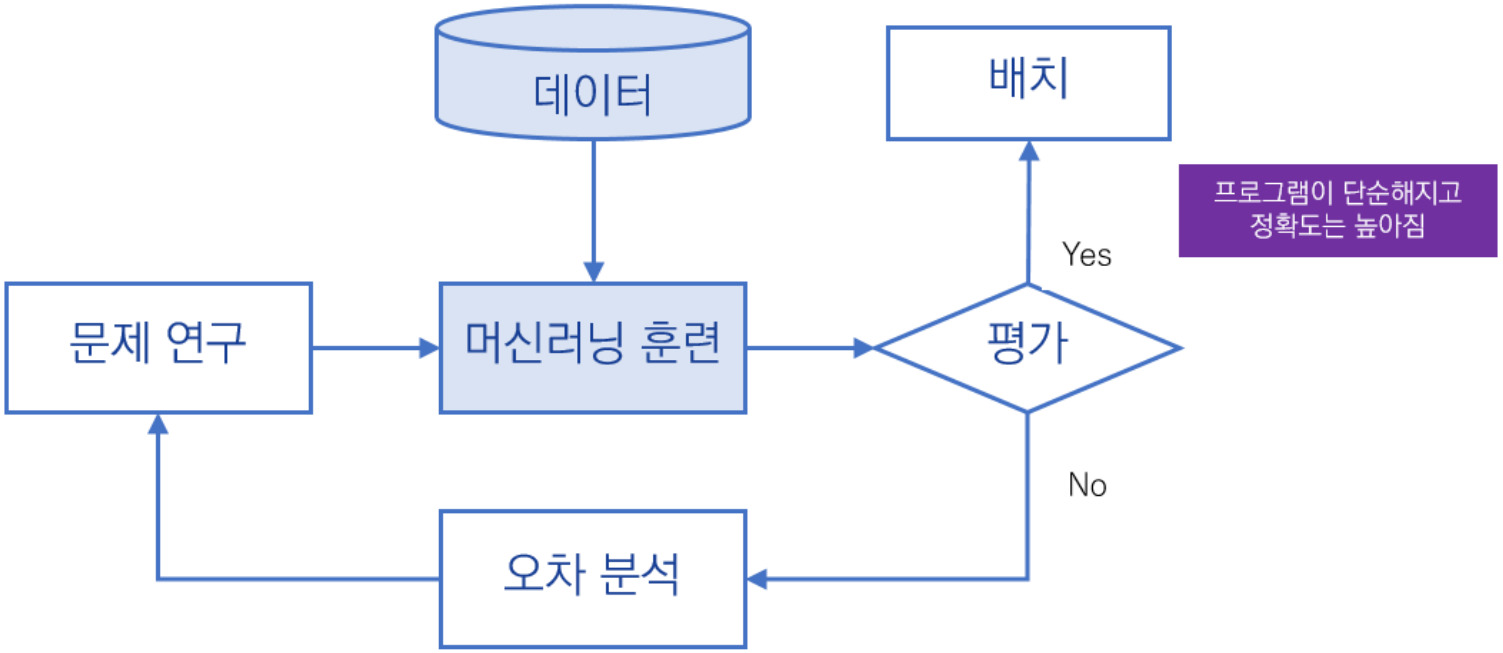
- 데이터로부터 학습하도록 컴퓨터를 프로그래밍하는 과학
- 명시적인 프로그래밍 없이 컴퓨터가 스스로 학습하는 능력을 갖게 하는 연구 분야(feat. Arthur Samuel)
- 과거 경험에서 학습을 통해 얻은 지식을 미래의 결정에 이용하는 컴퓨터 과학의 한 분야
- 관측된 패턴을 일반화하거나 주어진 샘플을 통해 새로운 규칙을 생성하는 목표를 가짐



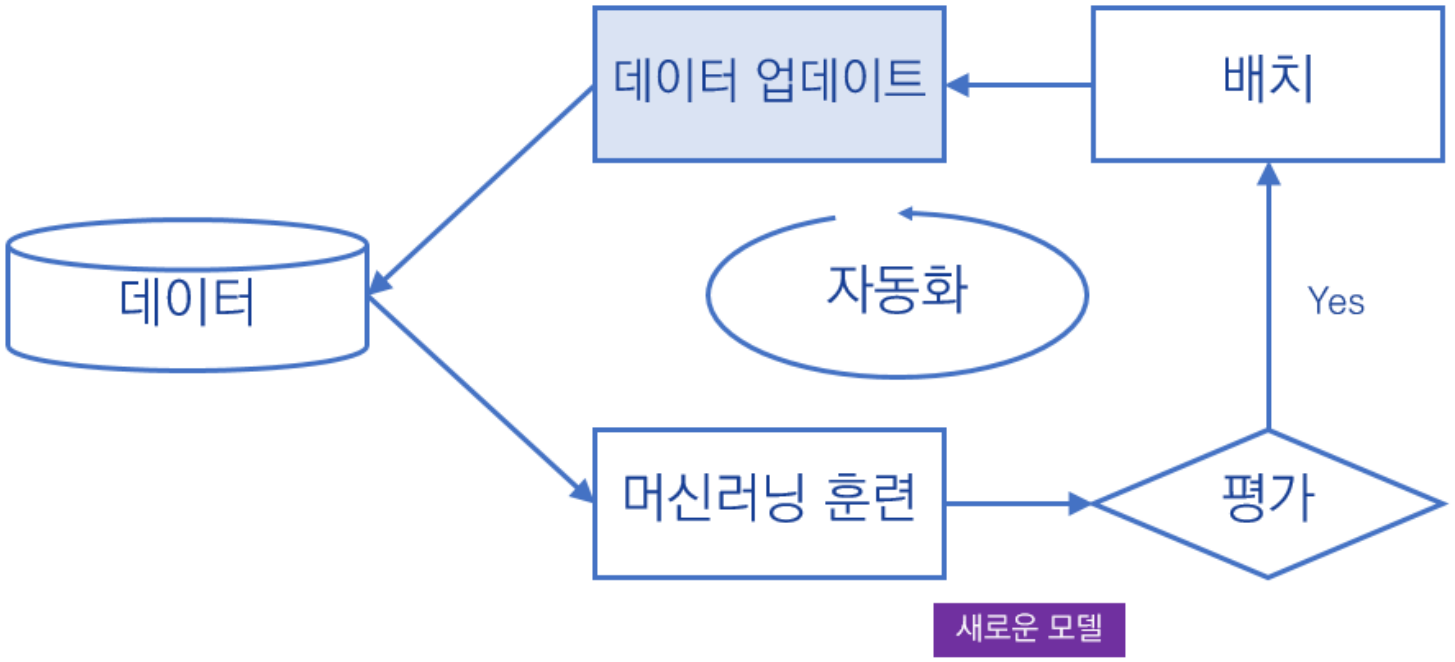
전통적인 접근 방법



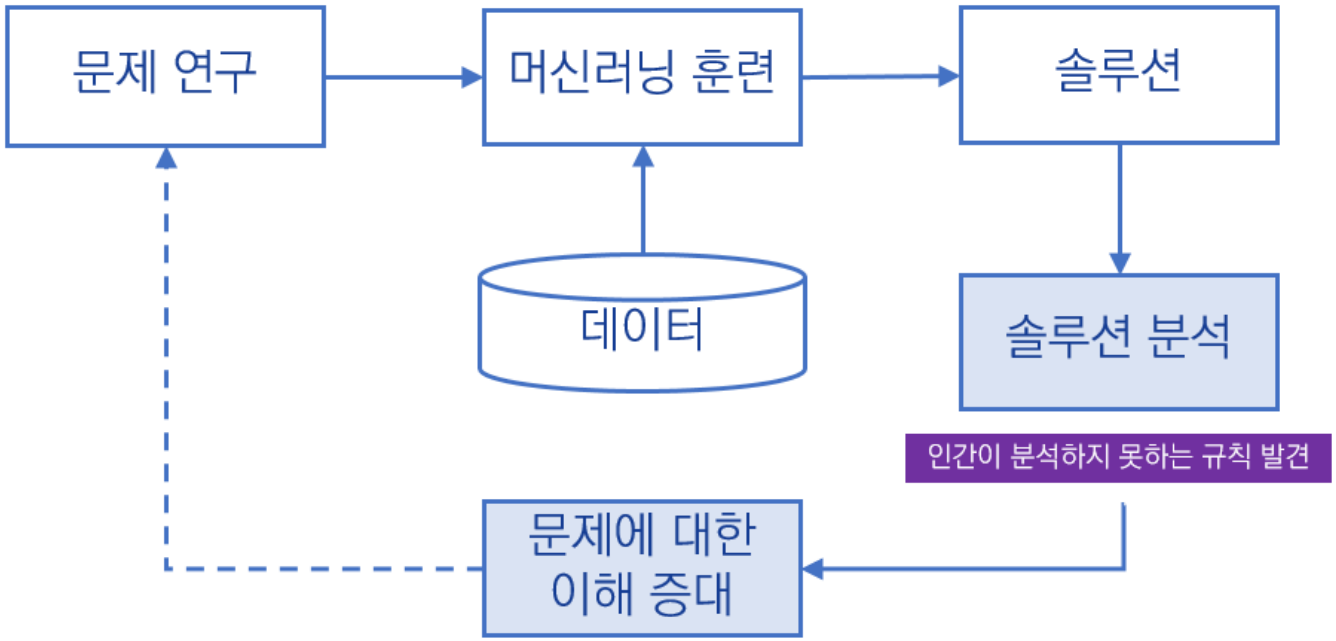
머신러닝 접근 방법



머신러닝 접근 방법 – 변화에 대한 적응 능력

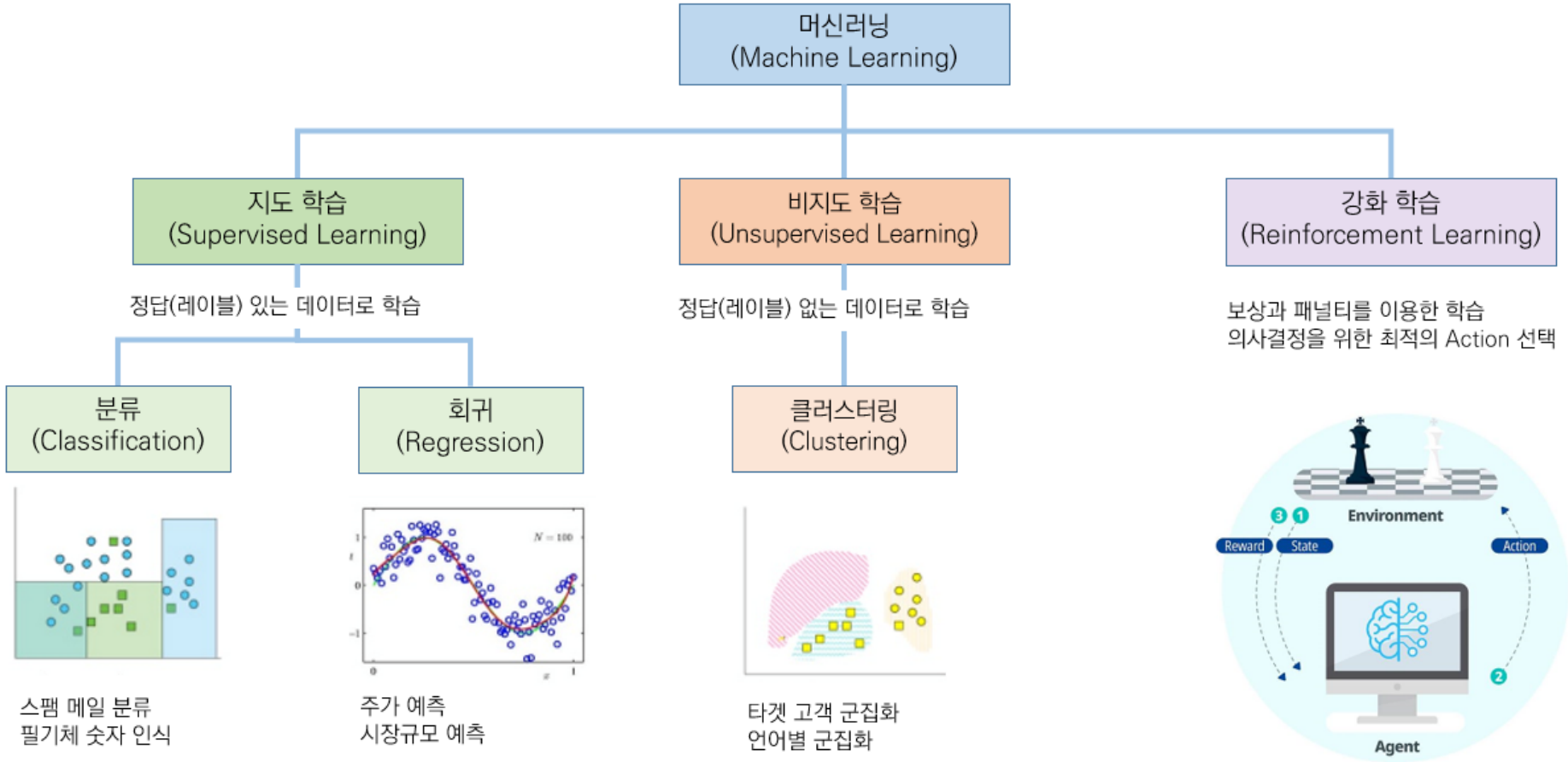


머신러닝 접근 방법 – 새로운 패턴 발견



머신러닝의 분류

- 데이터 학습 과정에서 정답(레이블) 유무에 따라 지도학습과 비지도 학습으로 나눌 수 있으며,
- 행동심리학에 기반하여 상태와 행동에 따른 보상을 통한 학습 방법인 강화학습 등이 있음



머신러닝의 분류 – 지도 학습

- 학습 데이터가 입력(특징 행렬)과 출력(대상 벡터) 쌍으로 제공됨 → "레이블 데이터"
- 학습 목표는 입력 특징 행렬과 출력 대산 벡터를 매핑시키는 규칙을 찾는 것임
- 입력 특징 행렬에 대해 출력 대상 벡터가 알려져 있으므로 '지도'라 부름

- 지도 학습 적용가능 대표적인 문제들

회 귀

- 연속형 수치 데이터 예측
- 집 값, 중고차 가격, 주가 예측 등

분 류

- 범주형 데이터인 클래스 레이블 예측
- 스팸 메일 필터, 긍정/부정의 감정분석, 채무불이행 예측 등

- 대표적인 지도 학습 알고리즘

k-최근접 이웃, 선형 회귀, 로지스틱 회귀, 서포트 벡터 머신, 의사결정 트리, 랜덤 포레스트, 신경망 등

머신러닝의 분류 – 비지도 학습

- 학습 데이터로 특징 행렬만 제공 → “레이블 없는 데이터”
- 입력 특징 행렬에 대한 출력 대상 벡터가 없으므로 ‘비지도’ 라 부름

- 비지도 학습 적용가능 대표적인 문제들

군 집

- 특징이 비슷한 것들끼리 묶어 군을 만드는 것
- K-평균, 계층 군집 분석(HCA), 기대값 최대화 등

시각화와 차원축소

- 시각화 시 인간이 인지할 수 있는 차원(2차원 등)으로 축소하는 것
- 주성분분석(PCA), 커널 PCA, 지역적 선형 임베딩(LLE), t-SNE

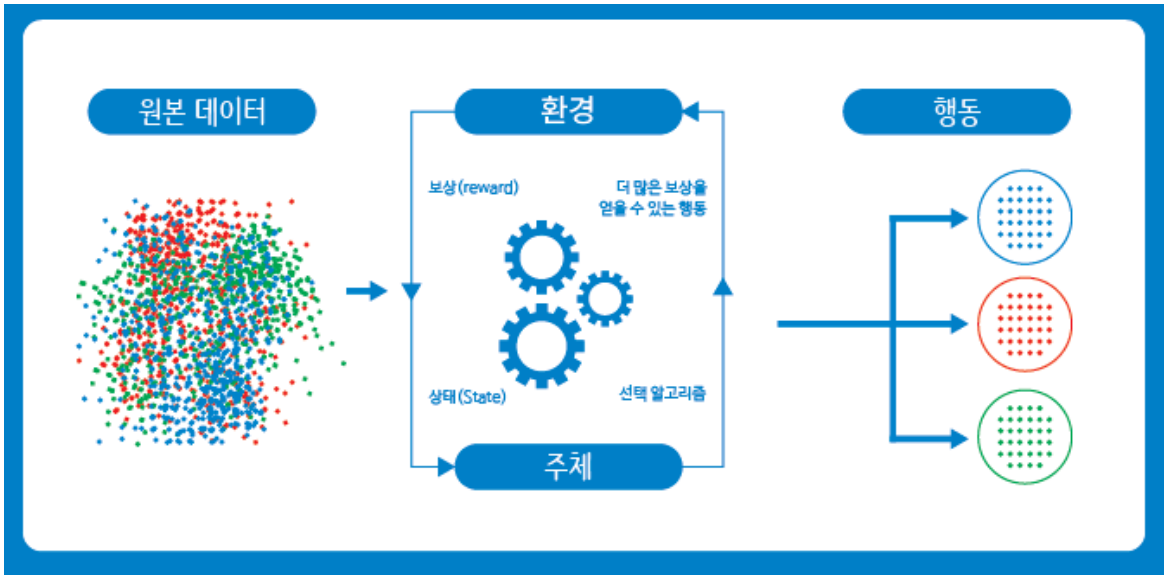
연관 규칙 학습

- 구매 경향성 규칙 발견 등(장바구니 분석)
- Apriori 알고리즘

- 추천시스템 유형도 존재함

머신러닝의 분류 – 강화 학습

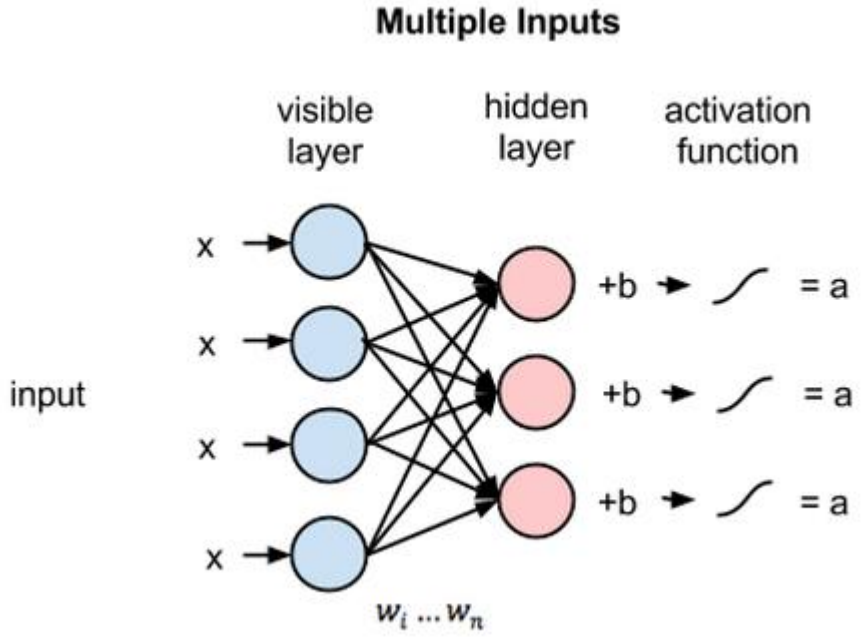
- 시스템이 어떤 목표를 달성하기 위해 동적인 조건에 적응하도록 함
- 학습하는 시스템(“에이전트”)이 환경 관찰 → 액션 실행 → “보상” 또는 “벌점”
- 시간이 경과하면서 가장 큰 보상을 얻기 위해 최상의 전략(“정책”)을 스스로 학습
- 자율주행자동차, 알파고 등에 활용



출처 : http://tcpschool.com/deep2018/deep2018_machine_reinforcement

머신러닝의 분류 – 준지도 학습

- 학습 데이터에 레이블이 일부만 있는 경우 활용
- 데이터 세트 전체에 대한 레이블링을 수행하는데 고비용이 발생
- 지도학습과 비지도학습의 조합으로 이루어짐
- 심층신뢰신경망(DBN)은 제한된 볼츠만 머신(RBM)과 같은 비지도학습에 기초



통계학 개념

- 통계학 이란?

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

- 기술 통계학 : 연속형 데이터 → 평균, 표준편차와 같은 자료 요약 **키, 나이, 가격 등**

  범주형 데이터 → 빈도, 백분율과 같은 자료 요약 **성별, 성씨 등**

- 추론 통계학 : 표본 샘플링을 통해 일부 자료를 수집해 전체 모집합에 대한 결론을 유추

  추론은 가설 검정, 수치의 특징 계산, 데이터 간의 상관관계 등을 통해 이루어짐

- 통계 모델링이란?

  데이터에 통계학을 적용하여 변수의 유의성을 분석함으로써 데이터의 숨겨진 특징을 찾아내는 것을 통계 모델링이라고 함

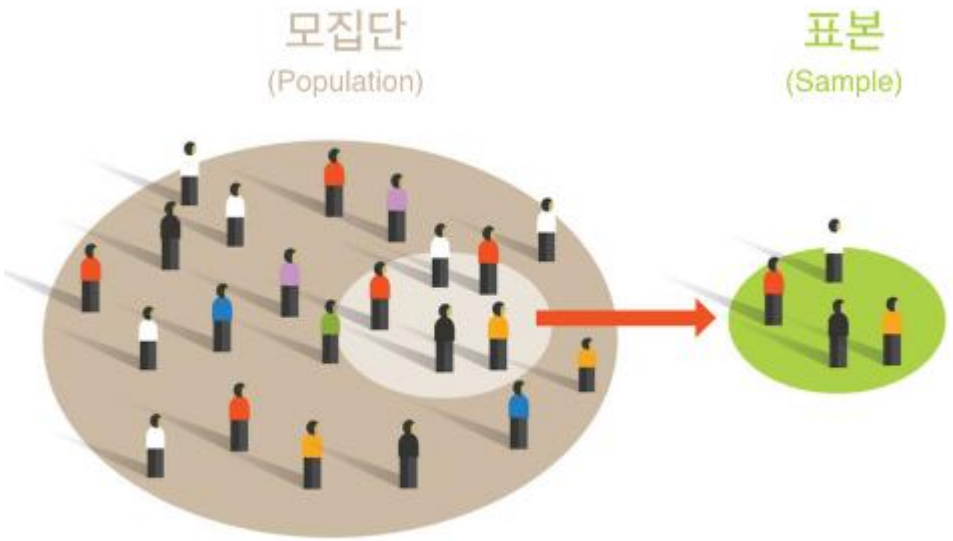
통계 모델 개념

- 통계 모델은 수학적 모델로 변수들로 이루어진 수학식을 계산해 실제 값을 추정하는 방법
- 통계 모델을 이루는 여러 가정은 확률 분포를 따름
- 통계 모델은 모든 변수가 만족해야 하는 기본 가정으로 시작하며,
  이 조건이 만족할 때만 모델의 성능이 통계학적으로 의미를 가짐

- 통계 모델의 구성
  - 데이터 : 변수, 관측점 및 값 등으로 구성되는데 컴퓨터가 인식할 수 있을 뿐만 아니라
    개념적으로도 일관성 있는 체계로 구성 필요
  - 함수 : 데이터에서 패턴을 추출할 수 있는 모형 훈련함수(training function)와 모형을 평가할 수 있는
    평가함수(evaluation function)로 구성
  - 공식 : 정해진 함수를 이용하여 변수들간의 관계를 표현

모집단과 표본

- 모집단 : 모든 관측값 또는 분석 대상의 전체 데이터를 의미
- 표본 : 모집단의 부분 집합으로, 분석 대상 중인 전체 데이터의 일부분

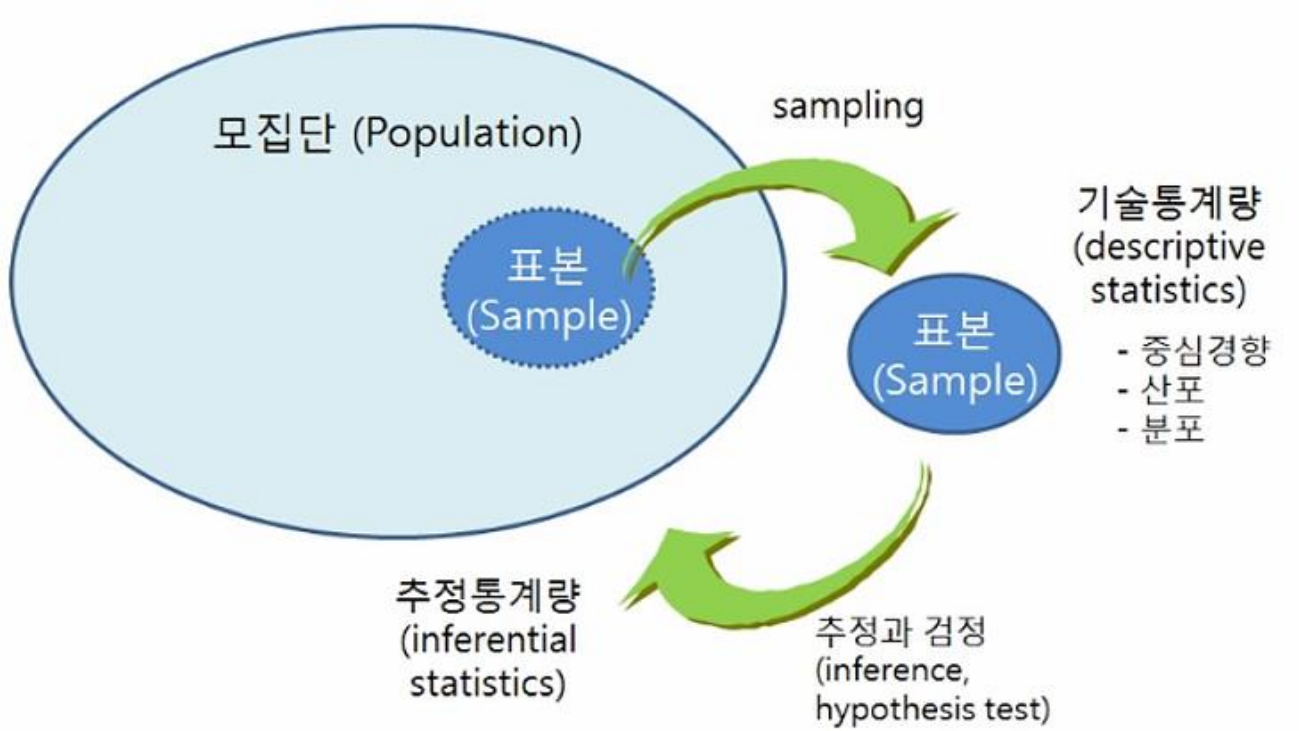


출처 : https://brunch.co.kr/@jaehyun-design/5

모수와 통계량

- 모수 : 모집단의 특징을 나타내는 수치값
- 통계량 : 표본의 특징을 나타내는 수치값으로, 모수 추정을 위해 사용

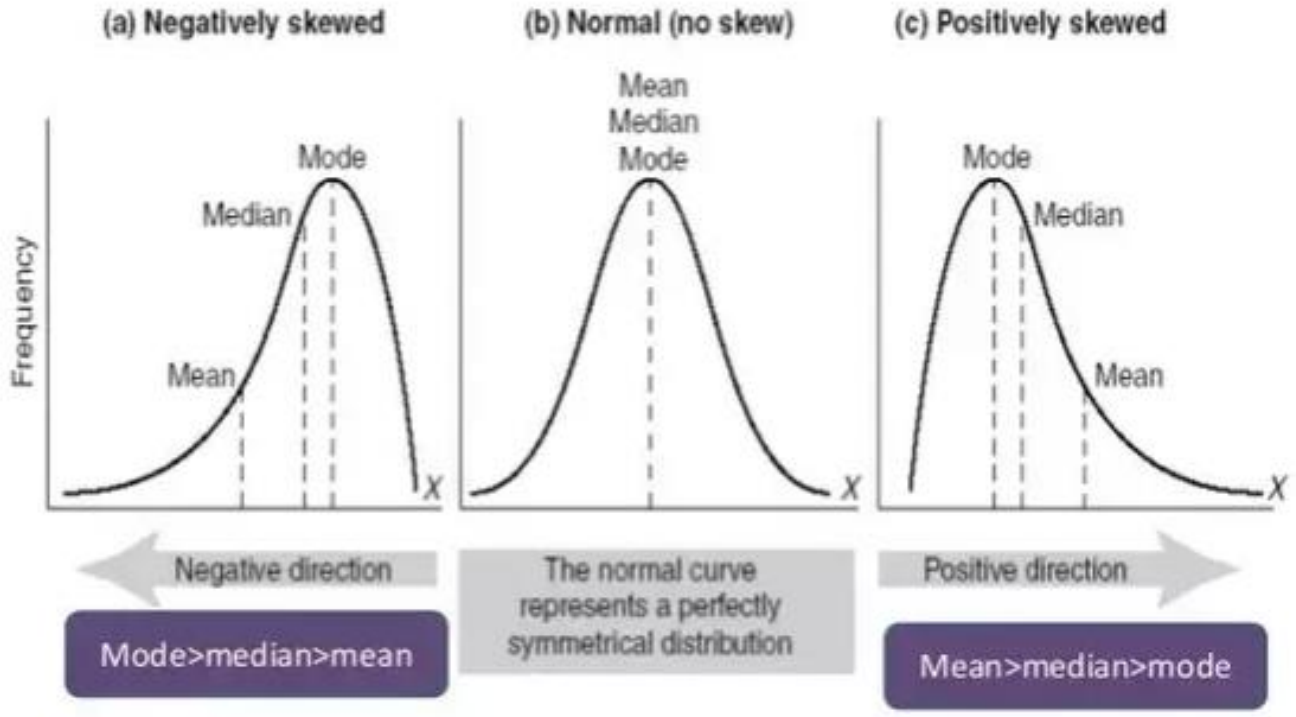
  평균, 중앙값, 최빈값, 분산 등과 같은 데이터를 대표하는 값



출처 : https://rlacksdid93.wixsite.com/930724/post/data-statistical-analysis-gicotonggyeryangyi-ihae

대푯값(representative value)

- 자료를 대표하는 값으로 평균, 중앙값, 최빈값, 백분위수, 사분위수, 절사평균 등
- 자료의 특징을 수치적으로 표현한 값



출처 : https://www.quora.com/How-is-the-gender-pay-gap-calculated-in-the-US

대푯값(representative value)

```
import numpy as np
from scipy import stats
```

```
np.random.seed(0)
data = np.random.randint(0, 100, 10000)
```

```
type(data)
```

```
numpy.ndarray
```

```
mean = np.mean(data); print('평균값 : {}'.format(mean))
median = np.median(data); print('중앙값 : {}'.format(median))
mode = stats.mode(data); print('최빈값 : {}({})'.format(mode[0][0], mode[1][0]))
```

```
평균값 : 49.1663
중앙값 : 49.0
최빈값 : 3(125)
```

numpy 라이브러리는 최빈값 관련 함수를 제공하지 않으므로, scipy 패키지의 status 모듈에 있는 mode() 함수를 사용

mode[0] : 최빈값
mode[1] : 최빈값의 빈도

변량(Variation)의 측정

- 변량의 측정 : 산포는 데이터의 변량을 의미하며, 데이터가 얼마나 중심으로부터 흩어져 있는가를 설명
- 분산 : 평균과의 거리를 제곱한 값의 평균
- 표준편차 : 분산의 제곱근
- 범위 : 최대값과 최소값의 차이
- 사분위수: 데이터를 4등분한 값으로 25% 값을 1사분위수(Q1), 50% 값을 2사분위수(Q2), 75% 값을 3사분위수(Q3)
- IQR(Interquartile Range) : Q1 과 Q3 의 차이

변량(Variation)의 측정

```
import numpy as np
from statistics import variance, stdev
```

```
np.random.seed(0)
points = np.random.randint(0, 100, 20)
```

```
var = variance(points); print('분산 :', var)
std = stdev(points); print('표준편차 :', np.round(std, 2))
range = np.max(data) - np.min(data); print('범위 :', range)
```

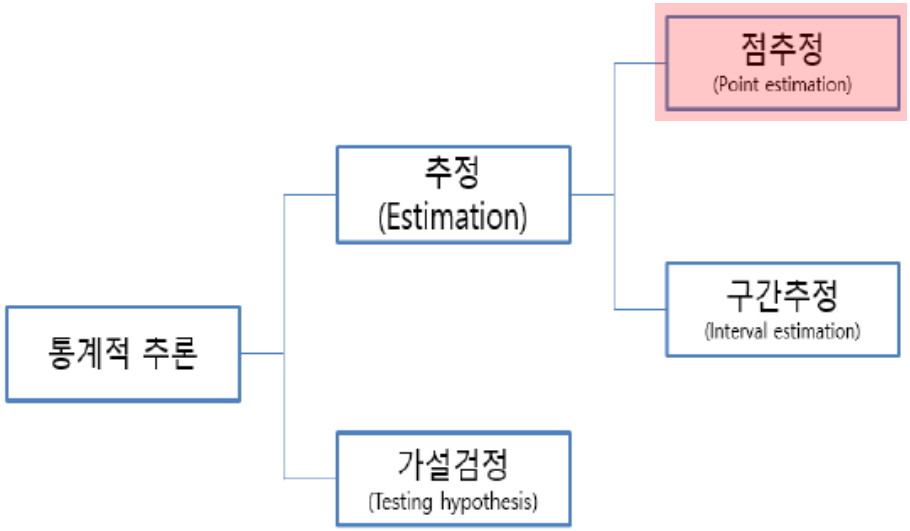
```
분산 : 662
표준편차 : 25.73
범위 : 79
```

```
print('[사분위수]')
for val in [0, 25, 50, 75, 100]:
    quantile = np.percentile(points, val)
    print('{}% : {}'.format(val, quantile))

q1, q3 = np.percentile(data, [25, 75])
print('IQR : ', q3 - q1)
```

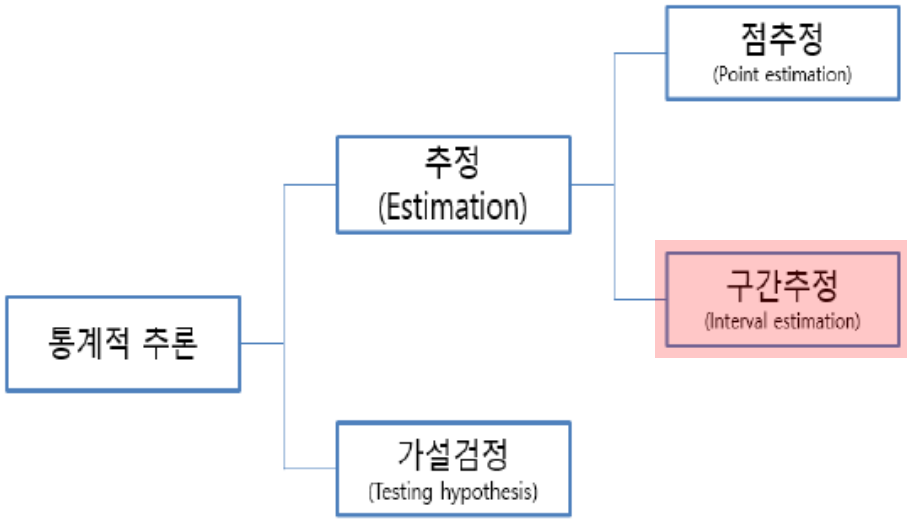
```
[사분위수]
0% : 9.0
25% : 42.75
50% : 64.5
75% : 84.0
100% : 88.0
IQR :  41.25
```

통계적 추론



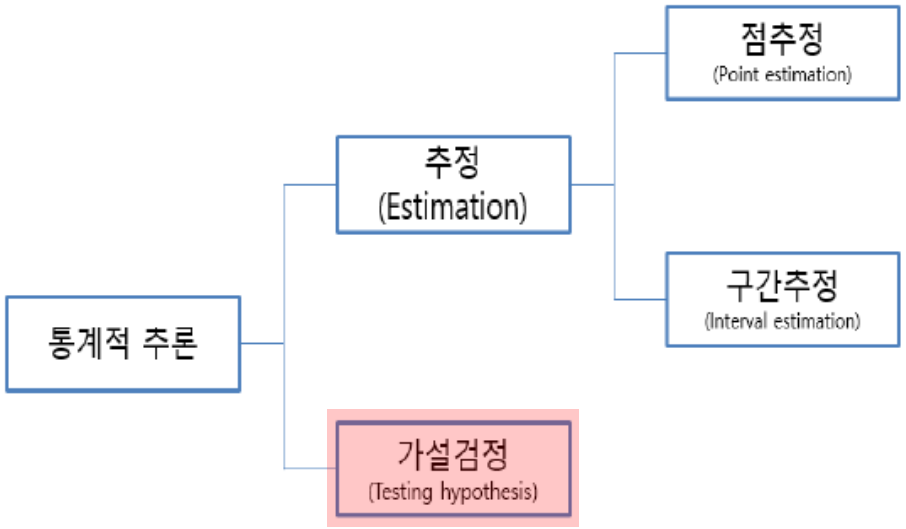
- 점추정(Point estimation)
  - 미지의 모수에 대해 표본의 통계량을 사용해서 특정 값으로 추정하는 과정
  - 모집단의 특성을 단일 값으로 추정하는 방법

통계적 추론



- 구간추정(Interval estimation)
  - 모수의 값이 포함될 것이라 생각되는 범위를 통해 모수를 추정
  - 모수의 구간 값을 계산해서 모수가 특정 구간에 포함 될 것을 확률로 분석
  - 신뢰수준으로 95%, 97% 등으로 확률로 표현

통계적 추론



- 가설검정(Testing hypothesis)
  - 모수에 대한 가설을 세우고 해당 가설의 채택|기각 여부 판단
  - 가설에 대한 검정을 통해서 기각|채택 여부 결정

가설검정(Testing Hypothesis)

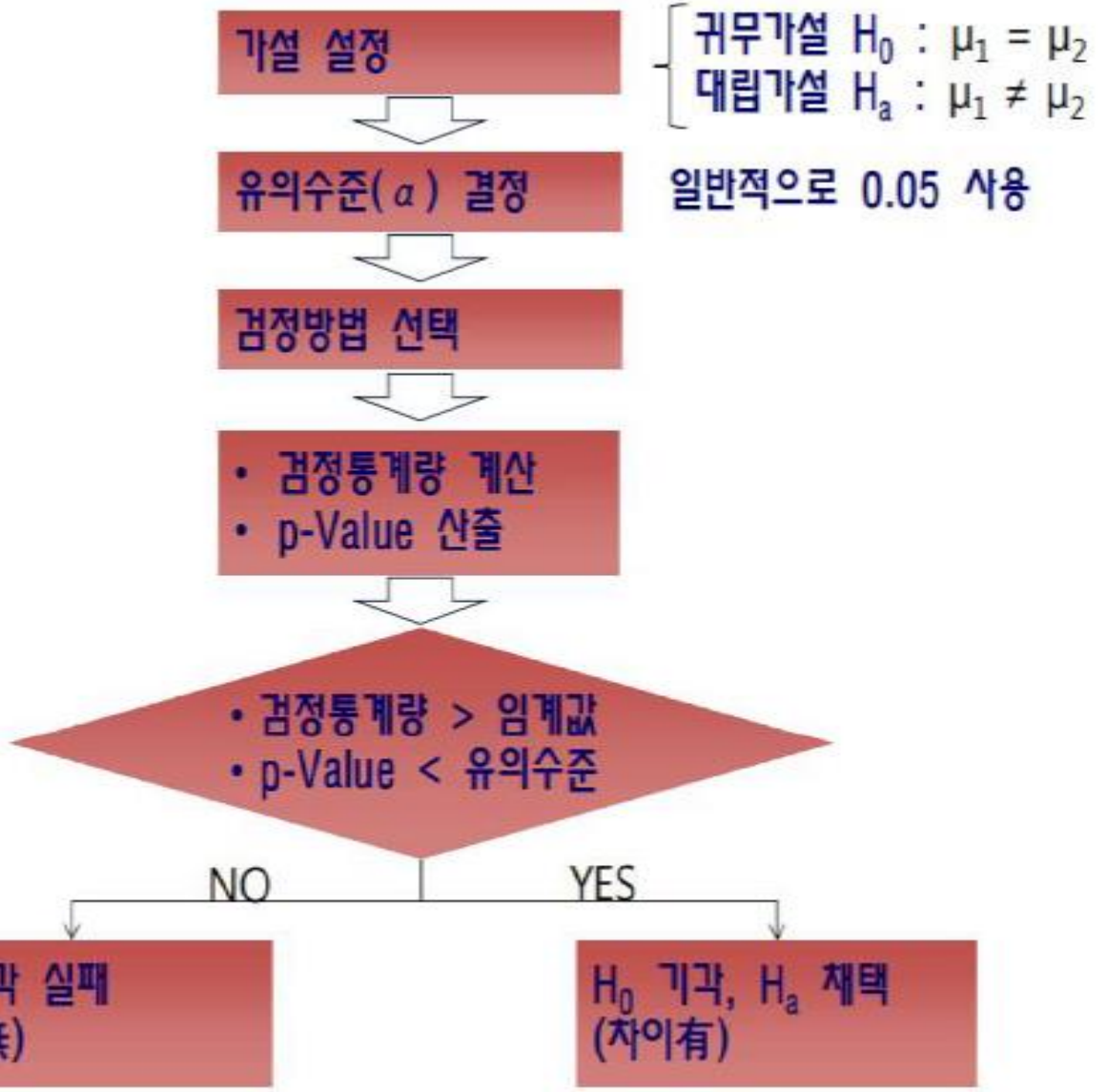
- 가설검정(Testing hypothesis)
  - 가설의 진실여부를 증명
  - 통계적 유의성을 검정하는 것으로 유의성 검정(Significance Test)라고도 함
  - 모수에서 표본을 사용하여 진실여부를 True 혹은 False로 판단
  - 귀무가설(H0)이 사실이라고 가정하고 검증 수행

가설검정(Testing Hypothesis)

- 귀무가설(Null Hypothesis, H0)
  - 기존에 알려진 사실, 일반적으로 진실이라고 믿고 있는 사실
  - 통계적 검정대상
  - 예) 모든 피고는 무죄이다. 제네시스는 연비가 10km이다.

- 대립가설(Alternative Hypothesis, H1)
  - 연구가설이라고도 함
  - 귀무가설과 대립하는 가설로 새로운 사실을 입증
  - 모수의 표본을 사용해서 검증
  - 예) 모든 피고는 유죄이다. 제니시스는 연비가 10km가 아니다.

가설검정(Testing Hypothesis) 절차



가설검정(Testing Hypothesis)

- p-value : 귀무가설이 옳다는 전제 하에 표본에서 실제로 관측된 통계값과 같거나 더 극단적인 통계값이 관측될 확률

"한 빵집에서 생산되는 식빵의 무게가 최소 200g이라고 주장할 경우, 표본 20개를 추출해 구한 평균 무게가 196g이고, 표준편차는 5.3g이었다면, 유의수준 5%(0.05)로 위의 주장을 기각할 수 있을까?"

- 귀무가설 : 모든 식빵의 무게는 200g 이상이다.

표본: $\bar{x} = 196,\ \sigma = 5.3,\ n = 20$ 표본 평균 / 표본 표준편차 / 표본의 크기

단일 표본에 대한 t-검정

```
import numpy as np
from scipy import stats
```

```
x_bar, mu, s, n = 196, 200, 5.3, 20
t_sample = (x_bar - mu) / (s / np.sqrt(float(n)))
print('검정통계량 :', np.round(t_sample, 2))

alpha = 0.05
t_alpha = stats.t.ppf(alpha, n-1)
print('임계값 :', np.round(t_alpha, 3))

p_value = stats.t.sf(np.abs(t_sample), n-1)
print('p-value :',np.round(p_value, 4))
```

```
검정통계량 : -3.38
임계값 : -1.729
p-value : 0.0016
```

임계값 : p-value=0.05일 때의 t 값

- p-value 0.0016 < 0.05이므로
→ 귀무가설을 기각

정규분포(Normal Distribution)

- 중심극한정리에 따르면 평균이 μ이고 분산이 σ^2(표준편차가 σ)인 모집단으로부터 가능한 모든 n개의 조합을 표본으로 추출하면 표본의 평균들은 정규분포에 접근함

$$X \sim N(\mu, \sigma^2)$$

"시험 점수가 정규분포를 따른다고 가정할 경우, 평균 점수가 56점이고 표준편차가 13.6인 경우, 75점 이상을 받은 학생들은 몇 %일까?"

정규분포를 표준정규분포(N(0, 1))로 변환

정규분포(Normal Distribution)

```
import numpy as np
from scipy import stats
```

```
x_bar, mu, sigma = 75, 56, 13.6

z = (x - mu) / sigma
print('z-score :', np.round(z, 2))

p_value = 1 - stats.norm.cdf(z)

print('학생이 {}점 이상의 점수일 확률 {}%'.format(x_bar, np.round(p_value * 100, 2)))
```

```
z-score : 1.4
학생이 75점 이상의 점수일 확률 8.12%
```

카이제곱(Chi-Square) 독립성 검정

- 카이제곱 독립성 검정

  범주형 데이터의 통계 분석에 가장 보편적으로 사용되는 검정

  2개의 범주형 변수 사이에 통계적 상관성이 존재하는지를 판단함

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e^i} \qquad o_i = \text{관측값}, \; e_i = \text{기대값}$$

"흡연이 운동에 영향을 미칠까?"

카이제곱(Chi-Square) 독립성 검정

```
import numpy as np
import pandas as pd
from scipy import stats
```

```
np.random.seed(0)
```

```
smoke = ['안함', '가끔', '매일', '심함']
exercise = ['안함', '가끔', '매일']
```

```
data = {'smoke' : np.random.choice(smoke, size=500),
        'exercise' : np.random.choice(exercise, size=500)}
```

```
df = pd.DataFrame(data, columns=['smoke', 'exercise'])
df.head()
```

	smoke	exercise
0	안함	안함
1	심함	안함
2	가끔	매일
3	안함	가끔
4	심함	가끔

카이제곱(Chi-Square) 독립성 검정

```
xtab = pd.crosstab(df.smoke, df.exercise)
xtab
```

카이제곱 검정을 위한 분할표 생성

smoke \ exercise	가끔	매일	안함
가끔	38	39	45
매일	34	44	33
심함	45	44	51
안함	47	41	39

```
contg = stats.chi2_contingency(observed = xtab)
p_value = np.round(contg[1], 3)
print('p-value :', p_value)
```

stats.chi2_contingency() : 카이제곱검정
(범주형 변수로 구성된 두 개의 집단에 사용)

```
p-value : 0.668
```

ANOVA 분산 분석

- ANOVA 분산 분석:

  Analysis of Variance

  모집단이 셋 이상인 경우, 이들의 평균이 서로 동일한가를 검정 모집단이 2개인 경우 t-검정을 수행

  귀무가설 : 모든 모집단의 평균은 동일하다.

  대립가설 : 적어도 하나의 모집단은 평균이 다르다.

"10명의 환자를 대상으로 A, B, C 세 가지 수면제 약효(수면시간)를 각각 테스트할 경우,
유의수준 0.05에서 A, B, C 수면제의 평균 수면시간은 동일한가?"

ANOVA 분산 분석

```
import numpy as np
import pandas as pd
from scipy import stats
```

```
np.random.seed(0)
```

```
data = (np.random.rand(30).round(2) * 10).reshape(-1, 3)
# data
```

```
df = pd.DataFrame(data, columns=['A', 'B', 'C'])
# print(df)
```

```
one_way_anova = stats.f_oneway(df.A, df.B, df.C)
print('통계량 : {}, p-value : {}'.format(np.round(one_way_anova[0], 2),
    np.round(one_way_anova[1], 3)))
```

```
통계량 : 0.34, p-value : 0.713
```

```
array([[2.6, 7.7, 4.6],
       [5.7, 0.2, 6.2],
       [6.1, 6.2, 9.4],
       [6.8, 3.6, 4.4],
       [7. , 0.6, 6.7],
       [6.7, 2.1, 1.3],
       [3.2, 3.6, 5.7],
       [4.4, 9.9, 1. ],
       [2.1, 1.6, 6.5],
       [2.5, 4.7, 2.4]])
```

	A	B	C
0	2.6	7.7	4.6
1	5.7	0.2	6.2
2	6.1	6.2	9.4
3	6.8	3.6	4.4
4	7.0	0.6	6.7
5	6.7	2.1	1.3
6	3.2	3.6	5.7
7	4.4	9.9	1.0
8	2.1	1.6	6.5
9	2.5	4.7	2.4

두 개이상의 요인에 대해 3집단 이상의 분석은 two-way ANOVA 분석 수행

ANOVA 분산 분석

ANOVA 분산 분석

37