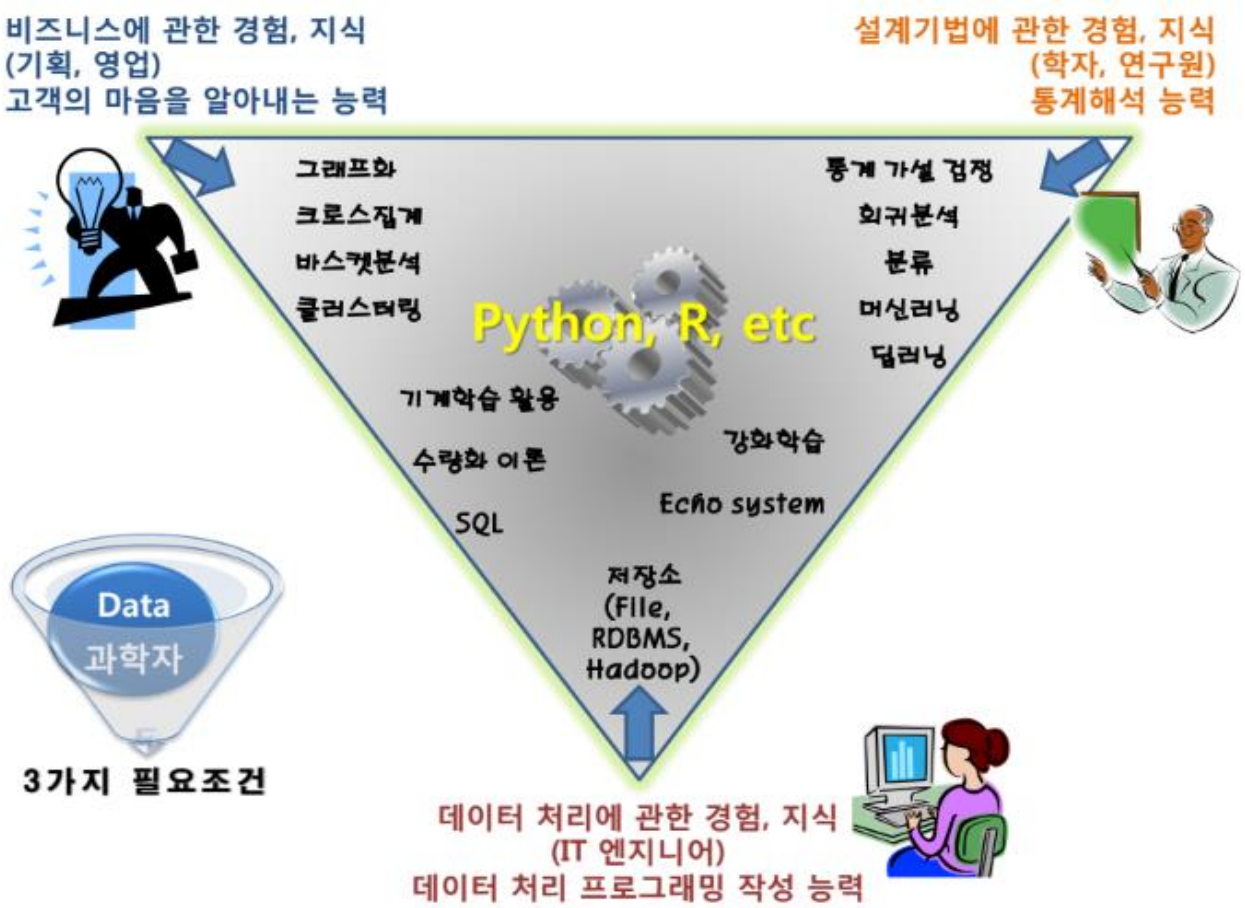
Python을 이용한 데이터 분석



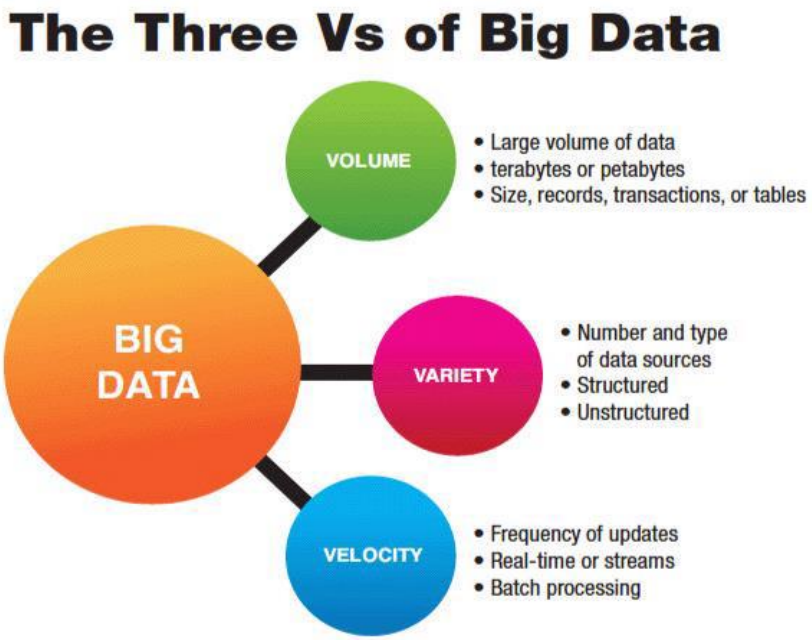
Park.yk

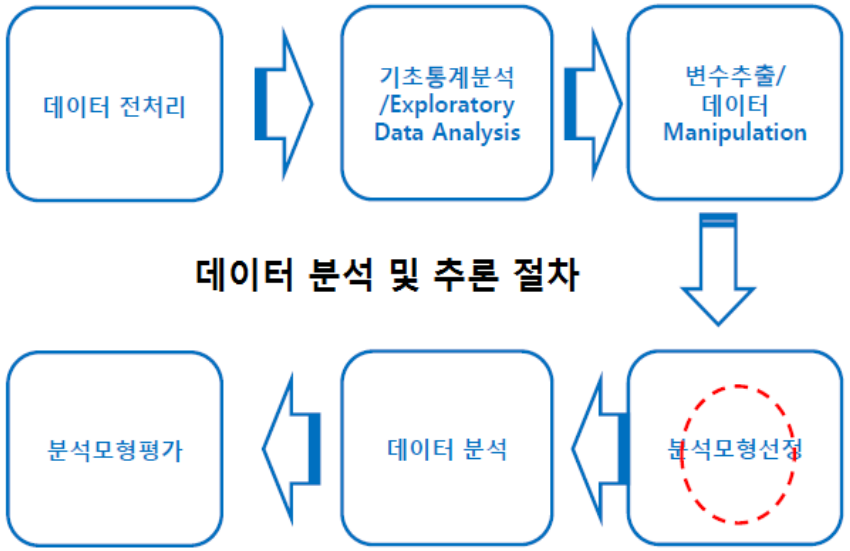
Big Data 정의

- 더 나은 의사결정, 시사점 발견 및 프로세스 최적화를 위해 사용되는 새로운 형태 의 정보처리가 필요한 대용량, 초고속 및 다양성의 특성을 가진 정보자산(Gartner)

- 일반적인 데이터베이스 소프트웨어 도구가 수집, 저장, 관리 및 분석하기 어려운 대규모의 Data.

- 3V를 가진 Data(Velocity, Volume, Variety)





통계학의 분류	
기술 통계학 (Descriptive statistics)	자료를 각 변수 별로 또는 관계되는 변수끼리 묶어서 요약. (평균, 분산, 표준편차, 사분위수, 중위수 등) 기술 통계는 추론 통계의 기초작업 수행을 위한 과정이라 할 수 있다.
추론 통계학 (Inferential statistics)	정리된 자료에 담긴 의미를 해석하여 미지의 세계에 대해 추론. (상관분석, 회귀분석, 분류, 인공신경망, 딥러닝 등)

* 데이터 분석을 위한 라이브러리 모음

참조 사이트 : pydata (http://pydata.org)

- numpy - 고속 연산
- scipy - 과학 분석 알고리즘
- pandas - 데이터 표현 및 처리
- matplotlib - 시각화 도구
- scikit-learn - Machine Learning, 추론 통계처리
- Tensorflow - Deep Learning
- Pytorch - Deep Learning

* Numpy *

- 데이터는 수 많은 숫자들로 이루어져 있다. 이 많은 숫자들을 효율적으로 계산하기 위해서는 관련된 데이터를 모두 하나의 변수에 넣고 처리해야 한다. 하나의 변수에 여러 개의 데이터를 넣는 방법으로 파이썬의 리스트를 사용할 수도 있지만 리스트는 속도가 느리고 메모리 소모가 크다. 더 적은 메모리로 더 빠르게 데이터를 처리하려면 배열을 사용하는 것이 좋다. 배열 사용을 위한 표준 패키지로 numpy가 있다.
- numpy는 수치 해석용 파이썬 패키지다. 다차원의 배열자료 구조인 ndarray 클래스를 지원하며, 벡터와 행렬을 사용하는 선형대수 계산에 주로 사용한다.

- Numpy는 과학적 계산을 위한 핵심 라이브러리다.
- 고성능 다차원 배열 객체와 배열과 함께 작동하는 도구들을 제공한다.
- 배열(ndarray) 처리를 위한 많은 함수를 제공한다.
- ndarray는 같은 타입을 가진 값들의 grid이며, 양의 정수 튜플로 인덱스 되어 있다.
- ndarray는 다차원 배열 객체로 파이썬의 리스트형 보다 처리속도가 빠르고 유연하다.

numpy

- np.array([1, 2, 3]) : 파이썬의 리스트 자료를 통해 rank가 1인 배열(1차원)을 생성한다.
- 슬라이싱(Slicing)이 가능하다. 모든 numpy 배열은 같은 타입을 갖는 요소의 grid이다.
- numpy는 배열을 만들 때 자료타입을 추측한다. 또한 배열 구성시에 명시적으로 자료타입을 선택적 인자로 포함하여 만들 수도 있다.
- **배열 연산(Array math)**

  기본적인 수학 함수는 배열에 요소별(elementwise)로 적용되고,

  연산자(+, -, *, /)나 혹은 numpy 모듈의 함수(add, subtract, multiply, divide)들을 사용할 수 있다.

  벡터화 연산을 하므로 반복문 없이 바로 배열에 대한 연산이 가능하다.

- 배열에 행/열 추가

- **Transpose(전치)**

배열에서 행과 열을 바꿀 수 있다. 이를 전치라 하며 데이터 모양이 바뀐 뷰를 반환 한다.

- **브로드캐스팅 (Broadcasting)은 크기가 다른 배열 간의 연산을 말한다.**

: 작은 배열과 큰 배열이 있을 때 작은 배열을 여러 번 반복해 큰 배열에 연산 수행 가능.

두 배열의 브로드캐스팅은 다음 규칙을 따른다.

1) 두 배열의 rank가 같지 않다면, 모양이 같은 길이를 가질 때까지 적은 rank 배열의 모양을 붙인다.
2) 두 배열이 그 차원에서 같은 크기를 갖거나, 차원에서 배열들 중 하나의 크기가 1이라면 두 배열은 차원에서 호환이 가능하다고 한다.
3) 브로드캐스팅 후에, 각 배열은 두 입력 배열들의 모양이 동일한 것처럼 행동한다.

Pandas

- 고수준의 자료구조(Series, DataFrame)와 빠르고 쉬운 데이터 분석용 자료구조 및 함수를 제공한다.
- NumPy의 고성능 배열 계산 기능과 스프레드시트
- SQL과 같은 RDMBS의 유연한 데이터 조작 기능을 갖고 있다.
- 세련된 인덱싱 기능으로 쉽게 데이터를 재배치하여 집계 등의 처리를 편리하게 한다.

- **Series**는 일련의 객체를 담을 수 있는 1차원 배열과 같은 자료구조로 색인을 갖는다.

```
obj = Series([3, 7, -5, 4])
# list, tuple type 가능. TypeError:'set' type is unordered

obj2 = Series([3, 7, -5, 4], index=['a', 'b', 'c', 'd'])
# 생성 시 색인을 지정
```

행렬은 수의 사각형 배열이다!

⇧

Scalar : 행렬이나 벡터의 각원소(실수)

Series : vector - 1차원 배열
DataFrame : marix - 2차원 배열
) 생성용 클래스

- 파이썬 dict type의 자료로 Series 객체를 생성")

```
names = {'mouse':12000, 'keyboard':25000, 'mornitor':'450000'}
```

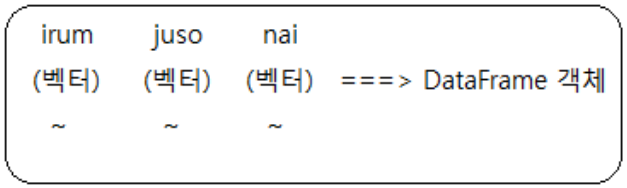
Pandas

DataFrame :

표 모양(2차원 형태 자료)의 자료구조로 여러 개의 칼럼을 갖는다. (Series가 모인 형태)
각 칼럼은 서로 다른 종류의 값을 기억할 수 있다.
같은 길이의 리스트에 담긴 dict type의 데이터를 이용해 DataFrame 객체 생성.

• 작성 예)

```
from pandas import DataFrame
data = {
   'irum':['홍길동', '한국인', '신기해', '공기밥', '한가해'],
   'juso':('역삼동', '신당동', '역삼동', '역삼동', '신사동'),
   'nai':[23, 25, 33, 30, 35],
}
```



- DataFrame의 칼럼은 사전형식이나 속성형식으로 접근 가능
- 슬라이싱
- 순서를 변경 및 칼럼에 값 대입으로 수정 가능
- DataFrame의 행 또는 열 삭제. 속성으로 axis=0 행, axis=1 열"
- 정렬. 재색인할 때 값을 보간하거나 채워 넣기
- 연산, 기술적 통계와 관련된 메소드
- 자료 합치기, group by, pivot
- 파일 읽기/저장

웹 문서 처리

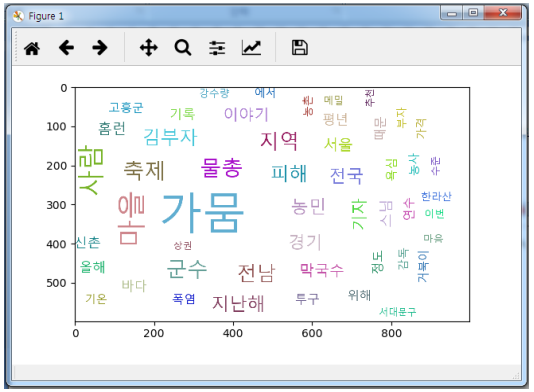
* BeautifulSoup으로 워드클라우드 그래프 그리기

  아래 모듈을 먼저 설치한다.

  ~> pip install pygame

  ~> pip install simplejson

  ~> pip install pytagcloud



* html 자료를 읽어 한글 형태소 분석하기 : konlpy

* Database 연동 후에 자료를 읽어 DataFrame 객체화 하기

SciPy

NumPy 기반으로 만들어졌다.

NumPy 배열에 작동하는 많은 수의 함수를 제공하며,

과학적이고 공학적인 응용의 다른 타입들에 유용하다.

Matplotlib

- http://matplotlib.org

- 플로팅 라이브러리로 matplotlib.pyplot 모듈을 사용하여 그래프 등의 시각화 가능.

- 그래프 종류 : line, scatter, contour(등고선), surface, bar, histogram, box, ...

- Figure

  모든 그림은 Figure라 부르는 matplotlib.figure.Figure 클래스 객체에 포함.

  내부 plot이 아닌 경우에는 Figure는 하나의 아이디 숫자와 window를 갖는다.

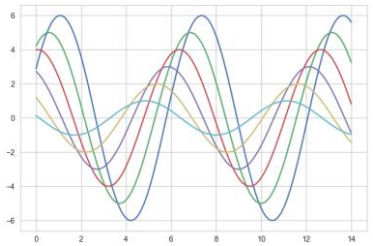
  figure()를 명시적으로 적으면 여러 개의 윈도우를 동시에 띄우게 된다.

- **matplotlib 의 기능 보충용 seaborn**

  : matplotlib에 seaborn을 사용하면 그래프를 더 멋지게 표현할 수 있다.

  : matplotlib 기반으로 다양한 색상 테마와 통계용 차트 등의 기능을 추가한 시각화 패키지다.

예)
```
import seaborn as sns
sns.set_style("whitegrid")
        ...
plt.show()
```



** 데이터 분석 **

- **데이터 분석의 목적**

데이터 분석이란 어떤 입력 데이터가 주어졌을 때 입력 데이터 간의 관계를 파악하거나 파악된 관계를 사용하여 원하는 출력 데이터를 만들어 내는 과정으로 볼 수 있다.

분석 목적에 따라 "예측(prediction)", "클러스터링(clustering)", "모사(approximation)" 등으로 나뉘어 진다.

참고로 예측 분석에 대해 살펴보자

- **예측(prediction)**은 데이터 분석 작업 중 가장 많이 사용되는 유형이다.

  이는 숫자, 문서, 이미지, 음성, 영상 등의 여러 입력 데이터가 주어지면 데이터 분석에 의한 예측된 결과를 출력한다.

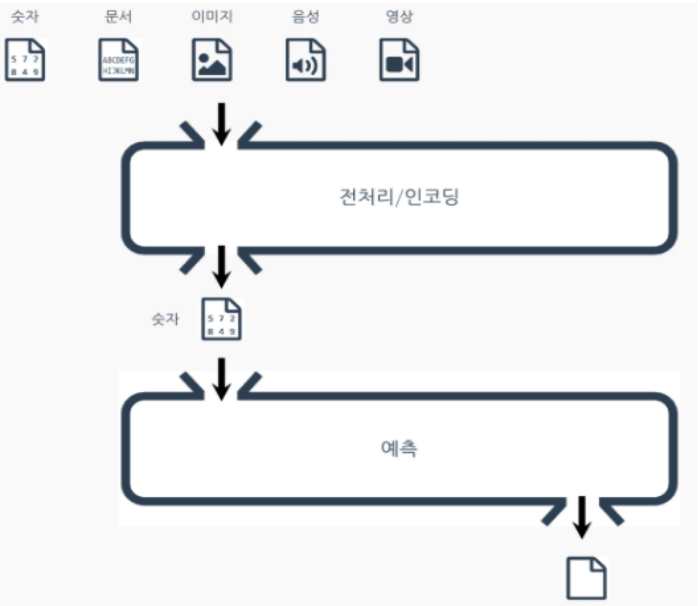
  예 1) 부동산의 위치, 주거 환경, 건축년도 등이 주어지면 해당 부동산의 가치를 추정한다.

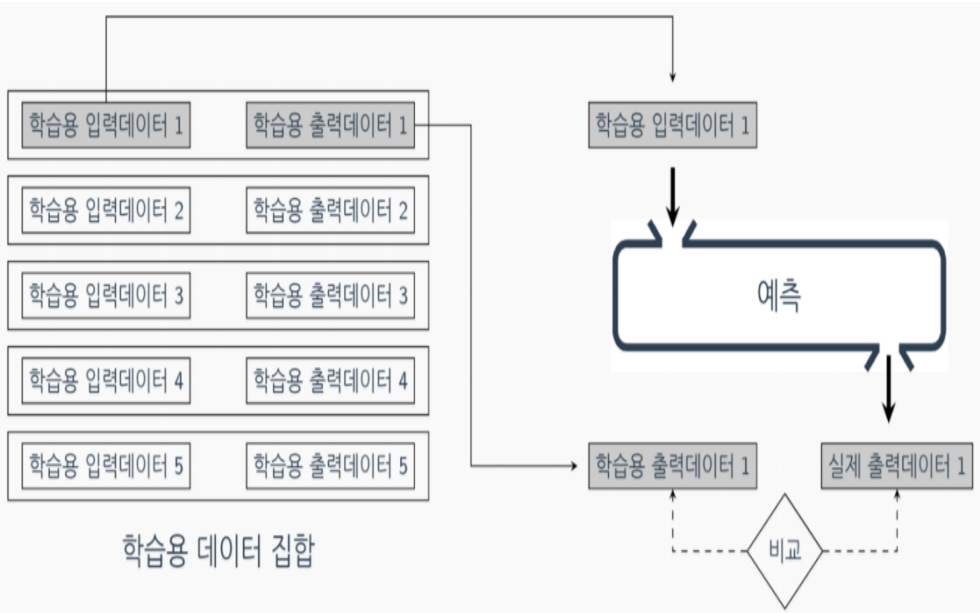
  예 2) 꽃잎의 길이와 너비 등 식물의 외형적 특징이 주어지면 해당 식물의 종을 알아낸다.

  예 3) 얼굴 사진이 주어지면 해당하는 사람의 이름을 출력한다.

** 데이터 분석 **

- 데이터 분석에서 말하는 예측이라는 용어는 시간상으로 미래의 의미를 포함하지는 않는다.
- 시계열 분석에서는 시간상으로 미래의 데이터를 예측하는 경우가 있는데 이 경우에는 미래 예측(forecasting) 이라는 용어를 사용한다.





머신러닝이란 프로그래머가 직접 수많은 규칙을 미리 정해주는 대신 프로그램이 데이터를 통해 스스로 학습하는 방법이다. 기존의 전통적인 프로그래밍에서는 사람이 전체적인 processing을 정해 주었으나 스펨메일 차단, 자율주행차 등의 작업을 처리하려면 다양한 상황에 대처할 수 있어야 한다.
머신러닝을 이용하면 explicit programming으로 해결할 수 없는 문제를 해결할 수 있다.
AI연구가 활발해짐에 따라 머신러닝 기법도 비약적으로 발전하고 있다.

** 데이터 분석 **

- **입력 데이터와 출력 데이터**

예측 문제에서는 데이터의 유형을 입력 데이터(input data)와 출력 데이터(output data)라는 두 가지 유형의 데이터로 분류할 수 있어야 한다.

- **입력 데이터는** 분석의 기반이 되는 데이터로 보통 알파벳 X로 표기한다.

  다른 말로 독립변수(independent variable), 특징(feature), 설명변수(explanatory variable) 등의 용어를 쓰기도 한다.

- **출력 데이터는** 추정하거나 예측하고자 하는 목적 데이터를 말한다.

  알파벳 Y로 표기하며 다른 말로 종속변수(dependent variable)라고 부른다.

  라벨(label) 또는 클래스(class)라고 하기도 한다.

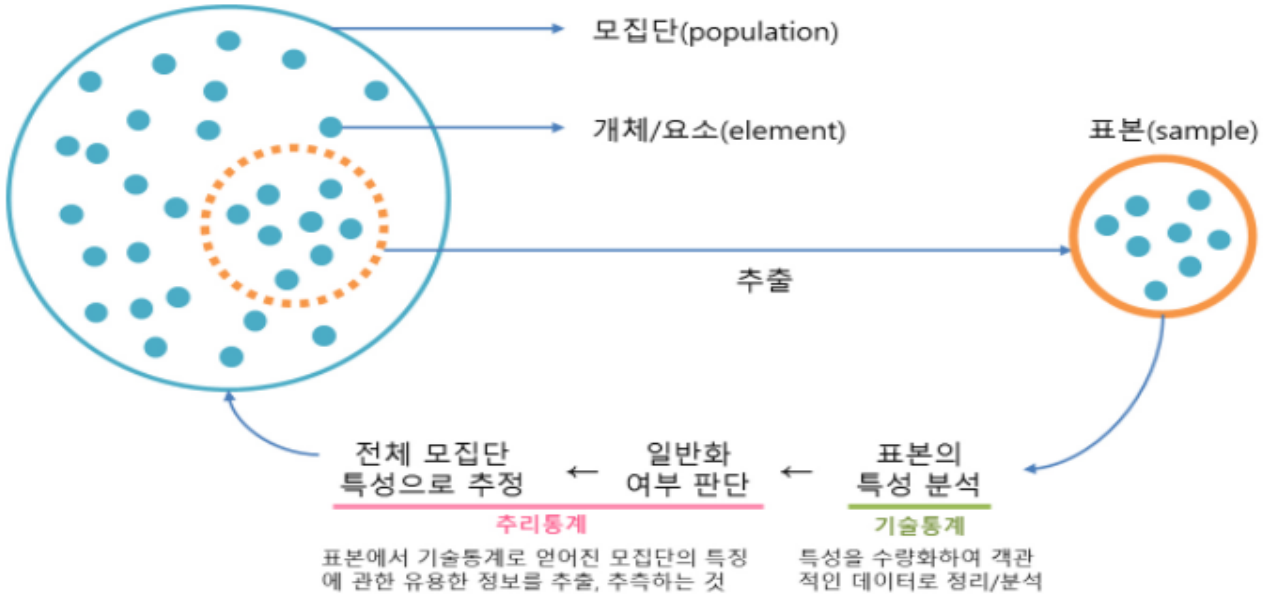
입력 데이터와 출력 데이터를 정확히 파악하는 것은 예측 문제를 구체화하는 첫 번째 단계로 예측 성능은 입출력 데이터의 숫자와 종류에 크게 의존하기 때문에 정확히 어떠한 값을 가지는 입력을 몇 개 사용하겠다는 문제 정의가 예측 문제를 해결하는데 중요한 부분이 될 수 있다.

기술통계 & 추리통계의 개념 정의

통계분석은 크게 **기술통계**(descriptive statistics)와 **추리통계**(inferential statistics)으로 나눌 수 있다.

기술통계	추리통계
수집한 데이터의 주요 특성을 분석 및 기술하는 통계방법 ex) 평균값 (mean),중위수 (median),최빈수 (mode),최대값, 최소값,범위 (range),분산 (variance),표준편차 (standard deviation) 등	수집한 데이터에서 표본(sample)을 추출, 특성을 파악하여 전체 데이터(모집단)의 특성으로 일반화할 수 있는 지 여부를 판단 모집단의 특성을 추정하는 것이 목적 - 간단히 표본을 기초로 향후의 일을 예측하는 것에 초점. ex)선거철.. 후보자의 지지도 조사
사례) H대학교 A학부의 최근 5년 간 4학년 학생들의 과목별 성적을 분석해서 학생들의 성적변화 추세치를 보려고 한다...	사례) B제품의 생산공장에서 라인별 제품의 불량률을 알아보기 위해 일정한 시간 간격으로 제품을 추출하여 분석하려 한다....

기술통계는 추리통계의 기초작업을 수행하기 위한 과정

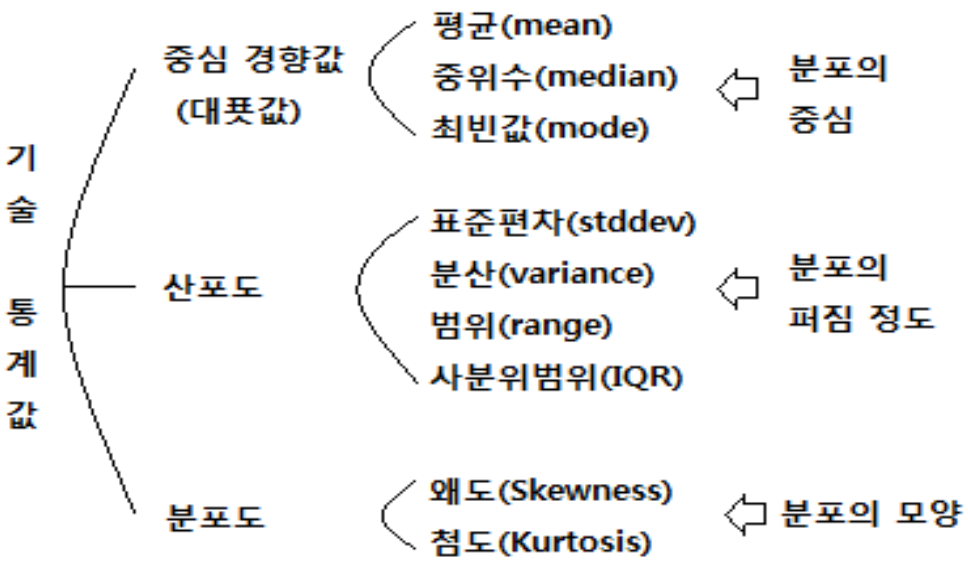


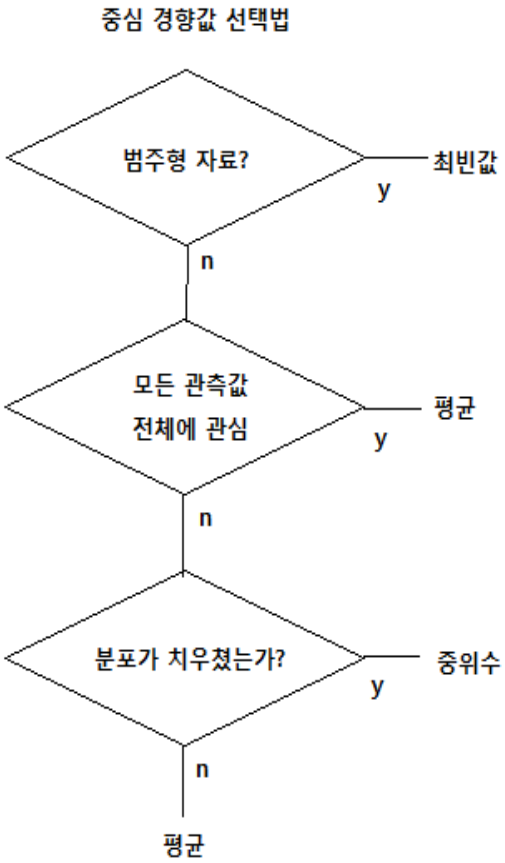
기술통계(Descriptive Statistics)

- 자료를 정리 및 요약하는 기초적인 통계
- 데이터 분석 전에 전체적인 데이터 분포의 이해와 통계적 수치 제공
- 추론통계의 기초자료로 많이 쓰인다.

기술통계량 유형 - 대표값, 산포도, 비대칭도 : 왜도, 첨도

기술 통계 분석 - 정보의 손실을 최대로 줄이면서 데이터를 효과적으로 요약할 수 있는 분석방법.





* 통계관련 기본 용어의 이해 *

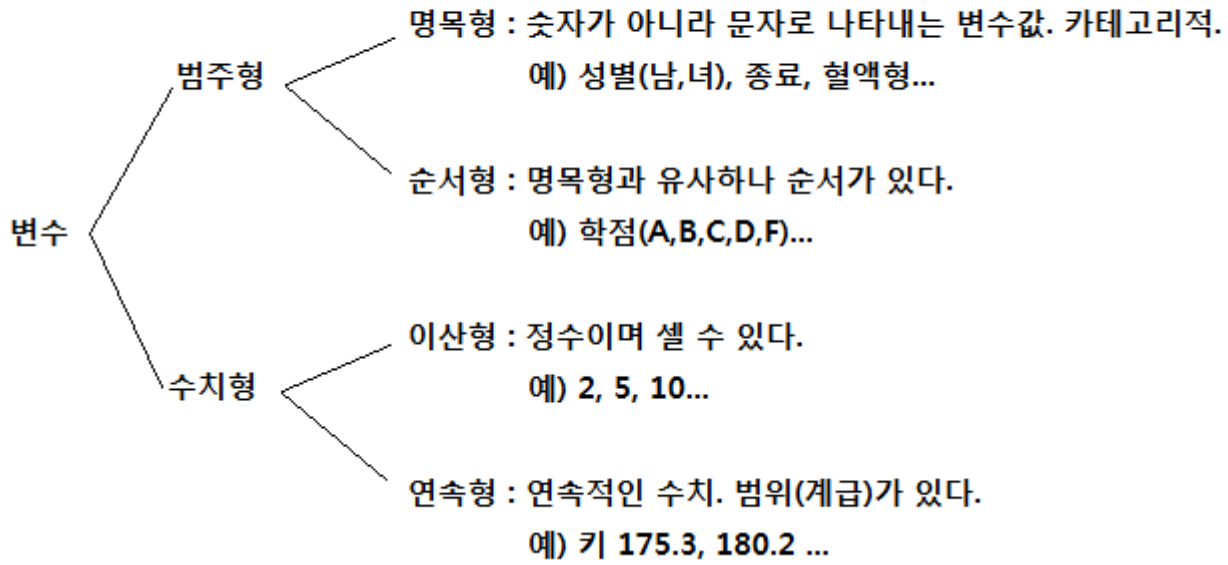
- 모집단의 통계수치를 모수라 하고, 표본의 통계수치를 통계량이라 한다.
- 오차는 평균으로부터의 치우침 이라 할 수 있다. 통계는 이러한 오차를 분석하고 관리하게 되는데 오차를 표현하는 대표적인 척도가 표준편차와 분산이다.
- 산포도는 변량이 흩어져있는 정도를 말한다. 변량들이 평균에 모여 있으면 산포도가 작다고 하고 변량들이 평균으로부터 떨어져 있으면 산포도가 크다고 한다. 산포도를 수치로 나타내는 방법으로는 분산과 표준편차가 주로 쓰인다.
- 대푯값은 자료들을 대표하는 값으로 쓰인다. 분산이나 표준편차 같은 산포도는 왜 필요할까? 대푯값이 자료들 모두를 반영하지 못하기 때문이다.
- 표준편차는 동일한 평균값을 갖는 둘 이상의 집단을 비교할 때 유용하게 활용된다. 예를 들어 어느 반 학생들의 기말시험 전체과목의 평균점수는 80점이라고 한다. 이 때에 A학생은 표준편차가 작고, B는 표준편차가 크다고 한다면 A학생의 수학점수는 80점 안팎으로 예측이 가능하지만 B학생의 수학점수는 예측하기가 어렵다. 그래서 통계적으로 결과를 추정할 때에는 모집단의 표준편차가 작을수록 보다 정확한 값을 예측할 수 있게 된다.

기술통계(Descriptive Statistics)

*** 척도(Scale) :**

- 자료가 수집될 때 관찰된 현상에 하나의 값을 할당시키기 위해 사용되는 측정의 수준
- "척도에 따른 분류"
  - 범주형(정성적 : 수량화가 불가 ex) 성별, 지역, 직업 등) - 명목형, 순서형(서열형)
  - 수치형(정량적 : 수량화가 가능 ex) 갯수, 나이, 키 등) - 등간, 비율

* 통계학에서의 데이터 종류



데이터에 따른 분석도구

추론 및 검정을 위한 데이터 분석 시에 분석 모델 선택에 영향을 주는 주요 구분 잣대로 사용되는 데이터 분석(추론 및 검정) 시 종속변수(반응변수)와 독립변수(설명변수)의 척도에 따라 분석 도구가 달라진다.

독립변수 (영향을 주는)	종속변수 (영향을 받는)	분석 방법
범주형	범주형	카이제곱 검정
범주형	연속형	T검정(범주형값 2개), ANOVA검정 - 범주형값 3개)
연속형	범주형	로지스틱 회귀
연속형	연속형	선형회귀, 구조 방정식

*** 추론통계 ***

통계학이란 논리적 사고와 객관적 사실을 바탕으로 일반적인 확률적 결정론에 의해서 인과관계를 규명하는 학문이다. 특히 연구목적에 의해 설정된 가설들에 대해서 분석결과가 어떤 결과를 뒷받침하고 있는지를 통계적 방법으로 검정할 수 있다.
기술통계가 수집된 자료의 특성을 쉽게 파악하기 위해 자료정리 및 요약을 하는 통계학 분야라고 한다면, 집단 간 차이분석은 모집단에서 추출한 표본정보를 통해 모집단의 다양한 특성을 추론하는 추론통계 분야다.

- **추정** : 모집단에서 추출한 표본에서 얻은 정보를 이용하여 모집단의 특성을 나타내는 값을 확률적으로 추정한다. 점추정, 구간추정
- **검정** : 유의수준과 표본의 검정 통계량을 비교하여 통계가설의 진위를 입증한다.

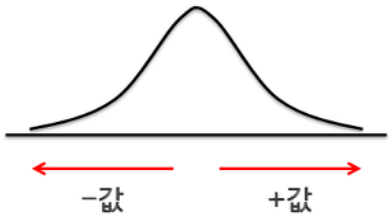
카이제곱분포

데이터들의 중심위치를 파악하는 대표적인 척도가 평균이다. 그리고 평균에서 데이터들이 흩어져 있는 정도, 즉 치우침을 표현하는 척도가 분산이다. (표준편차도 있다)

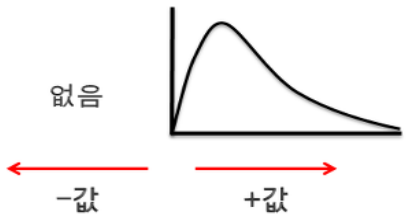
분산이 퍼져있는 모습을 분포로 만든 것이 카이제곱 분포다. 분산의 제곱된 값을 다루기 때문에 χ2분포라 불린다.

* 카이제곱분포 그래프의 특징
  - 확률변수는 연속확률 변수로서 항상 양(+)의 값을 갖는다.
  - 오른쪽 꼬리를 갖는 비대칭 분포다.
  - 자유도에 따라 모양이 다르다.
    자유도(df)가 커질수록 좌우대칭인 정규분포에 가까워진다.

<<정규분포>>



<<카이제곱분포>>



카이제곱 분포

카이제곱 분포 그래프는 0에 가까울수록 분포가 많고, 0에서 멀어질수록 분포가 감소한다. 이유는 데이터나 집단의 치우침은 어느 정도 크기인 경우가 많지, 치우침이 말도 안 되게 큰 경우는 별로 없기 때문이다.

예를 들어 성인 남자의 평균 키가 173cm라는 것은, 174, 169, 172 처럼 평균을 기준으로 치우침이 별로 크지 않은 사람이 많고, 상대적으로 198, 140 처럼 치우침이 아주 큰 사람은 적다. 그러므로 카이제곱 분포는 0에 가까울수록(치우침이 작을 경우) 분포가 많고, 0에서 멀어질수록(치우침이 클 경우) 분포가 감소하는 형태를 띠고 있다.

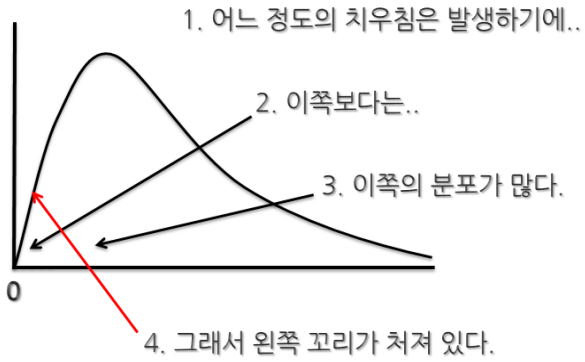
카이제곱 검정의 3가지 목적

1) 독립성 : 두 범주형 변수 간에 관련성이 있는지 여부를 알고자 할 때
2) 적합도 : 두 데이터가 특정한 분포에서 추출된 것인가 알고자 할 때
3) 동질성 : 두 개 이상의 다항분포가 동일한지 여부를 알고자 할 때

카이제곱 분포는 t분포와 마찬가지로 연속확률 분포이면서

표본분포로써 확률을 구할 때 사용하는 것이 아니라 가설검정을

할 때 주로 사용한다.



카이제곱분포

연구문제 : 성별에 따라 선호하는 커피브랜드의 차이가 있는가?

범주형 범주형
{성별}에 따라 {선호하는 커피브랜드}의 차이가 있는가?

범주형 자료에 따른 범주형 자료의 차이를 알아볼 때는? **카이제곱검정 활용**

성별/커피브랜드 모두 범주형 자료이고, 서로 어떤 영향을 미치는가 보기 위해, 카이제곱검정을 활용

▶ **카이제곱검정, 설문지 작성하기** 그렇다면 구체적으로 설문지 작성으로 넘어가 봅시다.
카이제곱을 사용하는 연구문제에서는 설문지를 어떻게 작성해야 할까?

설문지 구성 예시

1, 귀하의 성별은 무엇입니까?
① 남자 ② 여자

2. 선호하는 커피브랜드는 어디입니까?
① A사 ② B사 ③ C사

영향을 주고 받는 변수들 모두 범주형 자료로 구성되어 있는
카이제곱검정을 활용하기 위해서는,
범주형 자료를 얻을 수 있는 설문지 설계를 해야 한다는 것

카이제곱분포

▶ 연구문제 예시

그렇다면 또 어떤 연구문제에 카이검증을 적용할 수 있을까?

연구문제 예시

1. **20대와 60대**는 **여당과 야당을 지지하는 사람의 비율**에서 차이가 있을까?
2. **성별**에 따라서 **맥주를 좋아하는 사람과 소주를 좋아하는 사람의 비율**은 다를까?
3. **출신 지역**에 따라서 **특정 야구팀을 선호하는 사람의 비율**이 다를까?
4. **전공**에 따라서 **액션영화와 멜로영화를 좋아하는 사람의 비율**이 다를까?

- 독립변수 : **집단**
- 종속변수 : **어떤 특성의 비율**

위의 연구문제들은 모두 **집단**에 따라서, **어떤 특성의 비율 차이가 있는지**를 알아보는데 관심이 있다.

- 독립변수 : 집단
- 종속변수 : 어떤 특성의 비율

이처럼 **어떤 특성의 비율이 집단에 따라서 다른지에 대한 문제를 검증하고자 할때, 카이검증을 적용**할 수 있다

- 교차분석은 검정통계량으로 카이제곱을 주로 사용함(교차분석을 카이제곱 검정이라 함)
- 카이제곱은 주로 교차표를 작성하고, 두 변수 간의 독립성과 관련성을 분석한다.
- 카이제곱 검증 유형 분류 :
  일원카이제곱 검정(변인 단수 - 적합성),
  이원카이제곱 검정(변인 복수 - 독립성, 동질성)

일원 카이제곱 검정 실습

카이 제곱 검정은 goodness of fit(적합성) 검정이라고도 부른다.
SciPy stats 서브패키지의 chisquare 명령을 사용한다.

* 적합도 검정 실습 : 주사위를 던져서(60회) 관측도수 /기대도수가 아래와 같은 경우
적합한 주사위가 맞는가?'
* 적합성 가설 검정 예
  - 귀무가설 : 기대치와 관찰치는 차이가 없다. 예)주사위는 게임에 적합하다.
  - 대립가설 : 기대치와 관찰치는 차이가 있다. 예)주사위는 게임에 적합하지 않다.

주사위눈금	1	2	3	4	5	6
관측도수	4	6	17	16	8	9
기대도수	10	10	10	10	10	10

참고 – 가설 설정 방법
* 귀무가설 : 같다, 다르지 않다,
차이가 없다, 효과가 없다...
* 대립가설 : 같지 않다, 다르다,
차이가 있다, 효과가 있다...

이원카이제곱 실습

이원카이제곱

: 두 개 이상의 집단 또는 범주의 변인을 대상으로 동질성 or 독립성 검정.

: 유의확률에 의해서 집단 간에 '차이가 있는가? 없는가?' 로 가설을 검정한다.

동질성 검정 - 두 집단의 분포가 동일한가? 다른 분포인가? 를 검증하는 방법으로 두 집단 이상에서 각 범주(집단) 간의 비율이 서로 동일한가를 검정하게 된다. 두 개 이상의 범주형 자료가 동일한 분포를 갖는 모집단에서 추출된 것인지 검정하는 방법이다.

실습) 교육방법에 따른 만족도 분석 - 동질성 검정

동질성 분석

- 귀무가설 : 교육방법에 따른 만족도에 차이가 없다.
- 대립가설 : 교육방법에 따른 만족도에 차이가 있다.

집단 별 비율검정과 평균차이 검정

*** T검정(범주형 값 2개)**

: 비율 검정 - 빈도수에 대한 비율에 의미가 있다.

: 평균차이 검정 - 표본평균에 의미가 있다.

단일 표본 t-검정 (One-sample t-test)

- 단일 표본 t-검정은 정규 분포의 표본에 대해 기댓값을 조사하는 검정방법이다.
- SciPy의 stats 서브 패키지의 ttest_1samp 명령을 사용한다.
- 모수(평균)를 알고 있는 경우 sample의 평균과 모수(평균)와 여부를 검정

귀무가설 : 모수와 같다.

대립가설 : 모수와 다르다.

T검정, 차이검증이란? T검정은 **두 집단의 평균 점수를 비교**하고자 할 때, 실시하는 분석방법이다.

범주형 자료에 따른 연속형 자료의 차이를 볼 때, T검정과 ANOVA분석 (분산분석)을 사용할 수 있다.

영향을 주는 변수	영향을 받는 변수	통계분석방법
범주형 자료	범주형 자료	카이제곱 검정
	연속형 자료	T검정 분산분석

이 두 분석의 차이는 범주형 자료의 집단이 몇 개인가 이다.

범주형 자료의 집단이 두 개일 경우, **T검정**
범주형 자료의 집단이 세 개이상일 경우, **ANOVA분석 (분산분석)**

연구문제 : 성별에 따라 A사 커피브랜드의 만족도는 차이가 있는가? 통계분석 방법 : T검정

범주형 {성별}에 따라 연속형 {A사 커피 브랜드의 만족도}가 차이가 있는가?

[남자 / 여자] 두 집단

조건 1
범주형 자료에 따라 연속형 자료에 미치는 영향.

조건 2
남자 / 여자라는 범주형 자료의 두 집단.

} T-검정, T-Test

[조건 1] 성별이라는 범주형 자료에 따라, 만족도라는 연속형 자료를 확인하는 검증 과정이다.
[조건 2] 기준이 되는 성별이라는 범주형 자료가 남/녀로 두 집단이다.

▶T검정, 설문지 작성하기

그렇다면, T검정을 사용하는 연구문제에서는 설문지를 어떻게 구성해야 할까?

설문지 구성 예시

1, 귀하의 성별은 무엇입니까?
① 남자 ② 여자

2. A사 커피브랜드에 대해 전반적으로 얼마나 만족하십니까?
① 매우 불만족 ② 불만족 ③ 보통 ④ 만족 ⑤ 매우 만족

이처럼, **범주형 자료와 연속형 자료를 모두 얻을 수 있도록 설문지를 구성**해야 한다.

[범주형 자료] 비교하고자 하는 두 집단을 알아보기 위한 질문
[연속형 자료] 실질적으로 확인하고자 하는 변수를 알아보는 질문

T 검정

▶연구문제 예시

그렇다면, 또 어떤 연구문제에 T검정을 적용할 수 있을까?

연구문제 예시

1. **서울지역 고등학생과 부산지역 고등학생** 중에서 누구의 **수능점수**가 더 높을까?
2. **남학생과 여학생**은 **지능검사 점수**에서 차이가 있을까?
3. **무용학과 학생과 유아교육학과 학생** 중 어떤 학과 학생들의 **몸무게**가 더 높을까?
4. **중학생과 고등학생**의 **한 달에 받는 용돈**에는 차이가 있을까?

- 독립변수 : 집단
- 종속변수 : 어떤 특성의 평균값

위의 연구문제들은 모두 **두 집단 간에 어떤 특성의 평균 값에서 차이가 있는지**를 알아보는데 관심이 있다.

- 독립변수 : 집단
- 종속변수 : 어떤 특성의 평균값

이처럼 **두 집단 간에 어떤 특성의 평균값에서 차이가 있는지**에 대한 문제를 검증하고자 한다면 T검정, 차이검증을 적용할 수 있다.

두 집단 평균차이 검정 독립 표본 t-검정(Independent-two-sample t-test)

- 두 개의 독립적인 정규 분포에서 나온 두 개의 데이터 셋을 사용하여 두 정규 분포의 기댓값이 동일한지를 검사한다. SciPy stats 서브패키지의 ttest_ind 명령을 사용한다.
- 독립 표본 t-검정은 두 정규 분포의 분산 값이 같은 경우와 같지 않은 경우에 사용하는 검정 통계량이 다르기 때문에 equal_var 인수를 사용하여 이를 지정해 주어야 한다.

분산분석 중 세 집단 평균차이 검정(ANOVA : Analysis of Variance)

•선형회귀분석의 결과가 어느 정도의 성능을 가지는지는 단순히 잔차제곱합(RSS : Residuala Sum of Square)으로 평가할 수는 없다. 변수의 스케일이 달라지면 회귀분석과 상관없이 잔차제곱합도 같이 커지기 때문이다. ANOVA는 종속변수의 분산과 독립변수의 분산 간의 관계를 사용하여 선형회귀분석의 성능을 평가하고자 하는 방법이다. 분산분석은 서로 다른 두 개의 선형회귀분석의 성능 비교에 응용할 수 있으며, 독립변수가 카테고리 변수인 경우 각 카테고리 값에 따른 영향을 정량적으로 분석하는데도 사용된다.

• 독립변수가 복수(3개 이상)인 경우에는 각 독립변수에 대한 F검정 통계량을 구할 수 있다.

< F검정 통계량으로 가설검정 >

분산분석에서 신뢰수준 95%에서는 -1.96 ~ 1.96의 범위가 귀무가설의 채택역이다.

따라서 F검정 통계량이 채택역에 해당하지 않으면 귀무가설을 기각할 수 있다.

*** 분산분석에서 F검정 통계량과 유의수준 a(알파) 관계표**

F값(절대치)	유의수준a(양측검정 시)
F값(절대치) >= 2.58	a = 0.01 (의.생명 분야)
F값(절대치) >= 1.96	a = 0.05 (사회과학 분야)
F값(절대치) >= 1.645	a = 0.1 (일반 분야)

분산분석(ANOVA분석), 변량분석이란?

변량분석은 **둘 이상의 집단 간 평균 점수를 비교하고자 할 때 실시하는 분석방법**이다.
차이검증(T검정)으로는 두 집단 끼리만 비교할 수 있지만 변량분석을 이용하면 더 많은 집단끼리도 비교할 수 있다.

범주형 자료에 따른 연속형 자료의 차이를 볼 때, T검정과 ANOVA분석 (분산분석)을 사용할 수 있다.

이 두 분석의 차이는 범주형 자료의 집단이 몇 개인가 이다.

범주형 자료의 집단이 두 개일 경우, **T검정**
범주형 자료의 집단이 세 개 이상일 경우, **ANOVA분석 (분산분석)**

연구문제 : 직업(화이트칼라, 블루칼라, 주부, 학생)에 따라 A사 커피브랜드의 만족도는 차이가 있는가?
통계분석 방법 : ANOVA분석

범주형 {직업}에 따라 연속형 {A사 커피 브랜드의 만족도}가 차이가 있는가?

[화이트칼라 / 블루칼라 / 주부 / 학생] 네 집단

조건 1
범주형 자료에 따라 연속형 자료에 미치는 영향.

조건 2
기준이 되는 직업이 범주형 자료로 네 집단.

→ ANOVA 분석 (분산분석)

[조건 1] 화이트칼라, 블루칼라, 주부, 학생이라는 범주형 자료에 따라, 만족도라는 연속형 자료이다.
[조건 2] 기준이 되는 직업이 화이트칼라, 블루칼라, 주부, 학생이라는 범주형 자료로 네 집단이다.

▶ANOVA분석, 설문지 작성하기

그렇다면, ANOVA분석을 사용하는 연구문제에서는 설문지를 어떻게 만들어야 할까?

설문지 구성 예시

1, 귀하의 직업은 무엇입니까?
① 화이트칼라 ② 블루칼라 ③ 주부 ④ 학생

2. A사 커피브랜드에 대해 전반적으로 얼마나 만족하십니까?
① 매우 불만족 ② 불만족 ③ 보통 ④ 만족 ⑤ 매우 만족

[범주형 자료] 비교하고자 하는 세 개 이상의 집단을 알아보는 질문
[연속형 자료] 실질적으로 확인하고자 하는 변수를 알아보는 질문

이처럼, **범주형 자료와 연속형 자료를 모두 얻을 수 있는 설문지로 구성**해야 한다.
T검정과 유사하게 범주형 자료와 연속형 자료를 알아보지만, **ANOVA분석은 세 개 이상의 집단을 알아보는 질문으로 구성** 된다.

▶ANOVA분석, 결과 분석하기

연속형 자료인 만족도의 평균값을 도출한다. 데이터 수집을 바탕으로 얻은 결과값이 다음과 같다고 해 보자.

A사 커피브랜드 직업에 따른 만족도 평균

	평균
화이트칼라	3.14
블루칼라	2.97
주부	2.56
학생	2.47

- 명확한 차이가 있을 때에는 F 값 ↑, $p<0.05$
➔ **직업에 따라서 만족도는 유의미한 차이가 있다.**

통계적 검증을 하고 결과를 해석하는데 있어, **여기서부터 T검정과 조금** 다르다.
그 이유는 바로, **T검정**의 경우에는 **비교집단이 2개**이고 **분산분석**의 경우에는 **비교집단이 3개 이상**이기 때문이다.
비교집단이 2개인 경우에는 단순히 유의미한 차이가 있는지를 확인하면 된다.

하지만 비교집단이 3개 이상인 경우에는 **<1> 집단 간 유의미한 차이가 있는지를 확인**
<2> 각 집단끼리 어떤 차이가 있는지를 확인

집단이 2개이면 1번과 2번을 구분할 필요가 없게 되고, 집단이 3개 이상일 경우에는 1번과 2번 작업이 구분되는 것이다.

ANOVA

▶연구문제 예시

분산분석에서는 어떠한 연구문제에 적용할 수 있을까?

연구문제 예시

1. **20대, 30대, 40대** 간에 **패스트푸드에 대한 선호도**의 차이가 있을까?
2. **사무직, 기술직, 서비스직** 간에 **연봉**의 차이가 있을까?
3. **초등학생, 중학생, 고등학생, 대학생** 간에 **지능검사 점수**의 차이가 있을까?

- 독립변수 : 둘 이상의 집단
- 종속변수 : 어떤 특성의 평균값

위의 연구문제들은 모두 둘 이상의 집단 간에 어떤 특성의 평균값에서 차이가 있는지를 알아보는데 관심이 있다.

- **독립변수 : 둘 이상의 집단**
- **종속변수 : 어떤 특성의 평균값**

이처럼 둘 이상의 집단 간에 어떤 특성의 평균값에서 차이가 있는지에 대한 문제를 검증하고자 한다면, 분산분석, 변량분석을 적용할 수 있다.