



## ■ 데이터 전처리의 이해

데이터 전처리	데이터 분석을 위한 필수 과정으로 데이터를 정제한 뒤, 데이터 가공, 통합, 정리, 변환을 통해 데이터 분석 변수를 처리하는 등의 작업으로 데이터 분석 결과의 신뢰도를 높이기 위한 과정
---------	---

## ■ 데이터 정제

데이터 정제	결측값, 잡음, 이상값 등 데이터 오류의 원인을 분석 작업 전에 처리하는 것을 의미
결측값	분석대상에서 제외 또는 보완하여 처리 가능
이상값	삭제, 대체, 스케일링, 정규화 등의 방법으로 처리 가능

## ■ 데이터 결측값

결측값 유형	①비무작위(NMAR): 결측값에 영향 미친다. ②무작위(MAR): 연관은 있지만 결과에는 영향 미치지 않는다. ③완전 무작위(MCAR): 연관 없이 완전히 무관한 결측
결측값 대체 방법	①평균 대체: 대푯값으로 대체 ②단순 확률 대체: 단순 확률값으로 대체 ③보삽법: 비슷한 시기, 다른 해의 데이터를 참고한 평균값으로 대체 ④평가치 추정법: 맥락적/행렬식 자료를 고려하여 원래의 값 추정 ⑤다중 대체법: 결측치 추정을 통해 완성한 데이터셋을 이용하여 결측치 추정 ⑥완전 정보 최대우도법: 최대우도 바탕으로 가중평균 구성하여 대체

## ■ 데이터 이상값

이상값 검출	①분산: 정규분포 97.5% 외의 값 ②우도함수: 우도확률값 외의 값 ③근접 이웃 기반 이상치 탐지: 정상값 거리와 거리가 먼 값 ④밀도 기반: 상대적 밀도 값이 먼 값 ⑤군집: 특정 군집에 속하지 않는 값 ⑥사분위수: 양쪽 말단에 1.5분위수를 벗어나는 값
--------	---



## ■ 변수 선택

변수 선택	종속변수에 영향을 미칠 독립변수를 선택하는 과정, 선택적으로 변수를 적용하여 모델 성능 향상 가능	
선택적 변수	① 머신러닝 알고리즘 학습속도 향상	② 모델 해석 용이
선택의 이점	③ 모델 정확도 향상	④ 과적합 감소, 성능 향상

## ■ 단계적 변수 선택 방법

전진 선택법	가장 많은 영향을 줄 것 같은 변수부터 하나씩 추가(AIC 작은 것부터 추가)
후진 제거법	가장 적은 영향을 주는 변수부터 하나씩 제거(AIC 큰 것부터 제거)
단계적 방법	전진 선택법에 의한 유의한 변수 추가, 후진 선택법에 의한 유의성 낮은 변수 제거 작업 반복(변수 연속적 추가와 제거를 통해 AIC가 낮아지는 모델을 완성) ※AIC: 작을수록 좋은 데이터 모델이라고 할 수 있다.

## ■ 차원축소

차원축소	<ul style="list-style-type: none"> <li>변수의 수 증가로 인해 차원이 커지면서 데이터 모델링 성능 저하 문제가 발생하는 것을 '차원의 저주'라고 한다.</li> <li>공간은 증가하는 데 비해 데이터 수의 변화가 없는 경우 불필요한 정보와 공간으로 인해 모델링 성능의 저하를 유발할 수 있기 때문에 차원축소가 필요하다.</li> </ul>
PCA (주성분 분석)	<ul style="list-style-type: none"> <li>변수 간 상관관계를 파악하고 선형 연관성이 없는 저차원으로 축소하는 방법</li> <li>데이터 분산을 최대로 보존하는 축(PC1)을 찾고 PC1과 직교하면서, 두 번째로 분산이 최대인 축(PC2)을 찾는다. 이를 n회 반복하여 n만큼의 축을 찾는다.</li> </ul>
LDA (선형판별분석)	<ul style="list-style-type: none"> <li>지도학습을 통해 데이터 결정경계를 만들어 데이터 분류하는 것</li> <li>LDA의 2가지 가정               <ul style="list-style-type: none"> <li>①다변량정규분포를 따르는 데이터 분포여야 한다.</li> <li>②파라미터는 평균(벡터)과 공분산(행렬)이어야 한다.</li> </ul> </li> </ul>
t-SNE (t-분포 확률적 임베딩)	<ul style="list-style-type: none"> <li>고차원의 데이터 거리를 보존하며 그 관계를 저차원으로 축소하는 방법</li> <li>하나의 점(t1)을 선택하여 다른 점들 간의 거리를 측정한 뒤 이를 T 분포 그래프에 표현한 다음 t1을 중앙에 위치시키고 친밀도가 가까운 값끼리 그룹화</li> </ul>
SVD (특잇값 분해)	<p>행렬의 크기와 모양에 상관없이 적용할 수 있는 방법이다.</p> $(M = U \Sigma V^T)$

## ■ 변수 변환

범주형 데이터 변환	범주형 변수를 숫자로 변환 (남자:1, 여자:2)
연속형→범주형으로	연속형 데이터를 범주형으로 (10~19세: 10대)
비정형 데이터 변환	단어의 빈도수 등을 이용해서 정형화
더미 변수화	어떤 특징의 존재 여부를 1 또는 0으로 변환
스케일링	최소-최대 표준화, 정규화

## ■ 클래스 불균형

여러 클래스 중 데이터 양에 큰 차이가 있는 경우 클래스 불균형이 있다고 한다.

과소표집	소수 클래스의 데이터 수만큼 감소시킨다. 데이터 손실 우려.
과대표집	다수 클래스의 데이터 수만큼 증가시킨다. 과적합 문제 발생 가능.
SMOTE	주변값을 기준으로 소수 클래스의 데이터 수를 증가시켜 다수 클래스의 수와 동일해지게 한다.



## ■ EDA

### 탐색적 데이터 분석

데이터를 이해하고 의미 있는 관계를 찾아내기 위해 데이터의 통겅값과 분포 등을 시각화하고 분석하는 것으로, 줄여서 EDA라고 함.

## ■ EDA의 4R

저항성(Resistance) 강조

잔차(Residual) 계산

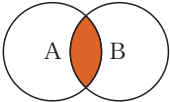
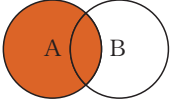
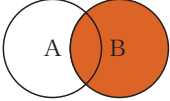
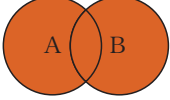
변수의 재표현(Re-expression)

그래프를 통한 현시성(Representation)

## ■ 데이터 탐색 개요

데이터 파악		분석 대상이 되는 데이터에 대해 기술한 설명서를 확인
데이터 탐색	데이터 조회하기	일부 관측치를 조회하여 개별 관측치들을 관찰
	데이터 구조 확인하기	데이터의 차원, 변수명, 변수 타입, 상위 관측치 등을 확인하여 데이터 구조를 간결하게 확인
	데이터 요약하기	데이터셋의 통계 요약량을 확인
	데이터 관계 맺기	현실의 데이터는 각각의 용도, 목적, 종류에 따라 각기 다른 테이블에 저장되어 있는데, JOIN 구문을 통한 병합이 가능

## ■ JOIN 종류

내부 조인 (INNER JOIN)	내부 조인을 사용하여 조인 조건을 만족하는 데이터를 가져올 수 있다.	
왼쪽 외부 조인 (LEFT OUTER JOIN)	왼쪽 외부 조인은 왼쪽 테이블을 기준 삼아 모든 행을 포함시키고 조인 조건에 부합하는 경우만 값을 가져와 결과에 포함한다.	
오른쪽 외부 조인 (RIGHT OUTER JOIN)	오른쪽 외부 조인은 오른쪽 테이블을 기준 삼아 모든 행을 포함시키고 조인 조건에 부합하는 경우만 값을 가져와 결과에 포함한다.	
완전 외부 조인 (FULL OUTER JOIN)	완전 외부 조인(FULL OUTER JOIN) : 중복되는 데이터는 삭제하고 왼쪽 외부 조인과 오른쪽 외부 조인 결과를 합집합으로 처리한 결과와 동일하다.	

\* 이미지 출처: <https://www.datasciencemadesimple.com/join-in-r-merge-in-r/>

## ■ 상관분석과 상관계수

상관분석	두 변수 사이에 선형적 관계를 가지고 있는지 분석하는 통계적 분석 방법
상관계수	두 변수의 선형 관계 정도를 나타내는 척도로, 피어슨 상관 계수가 대표적

## ■ 상관계수 해석

$\gamma$ 범위	관계
$0.7 \leq \gamma \leq 1$	강한 양의 상관관계가 있음
$0.3 \leq \gamma \leq 0.7$	뚜렷한, 보통의 양의 상관관계가 있음
$0 \leq \gamma \leq 0.3$	약한 양의 상관관계가 있음
$\gamma = 0$	선형 상관관계 없음
$-0.3 \leq \gamma \leq 0$	약한 음의 상관관계가 있음
$-0.7 \leq \gamma \leq -0.3$	뚜렷한, 보통의 음의 상관관계가 있음
$-1 \leq \gamma \leq -0.7$	강한 음의 상관관계가 있음



## ■ 피어슨 상관 계수 vs. 스피어만 상관 계수

피어슨	스피어만
$r_{XY} = \frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i^n (X_i - \bar{X})^2} \sqrt{\sum_i^n (Y_i - \bar{Y})^2}}$	$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$ $d_i = x_i - y_i$ <p><math>x_i</math>: 변수 x의 i번째 관측치의 순위  <math>y_i</math>: 변수 y의 i번째 관측치의 순위</p>
피어슨 상관계수 $r$ , 피어슨 적률상관계수(Pearson product-moment correlation coefficient)	스피어만 상관계수 $\rho$ (Spearman's rho), 스피어만 순위 상관계수(Spearman's rank correlation coefficient)
모수 검정	비모수 검정
연속형 변수	이산형, 순서형 변수
경영학 점수와 통계학 점수 사이에 연관성이 있는가?	경영학 과목 석차와 통계학 과목 석차 사이에 연관성이 있는가?

## ■ 기초통계량 추출 및 이해

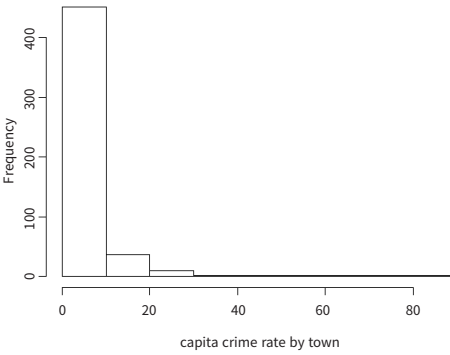
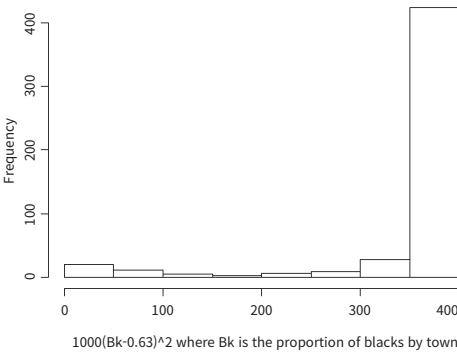
- **중심경향치**: 단일 값으로 전체 데이터를 대표할 수 있게 중앙에 위치한 데이터를 표현

평균	주어진 모든 데이터의 값을 더해 총합을 구하고 이를 데이터의 개수로 나눈 것 $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
중앙값	주어진 데이터의 값들을 오름차순 정렬했을 때 중간에 있는 값 <ul style="list-style-type: none"> <li>■ 데이터의 개수가 홀수인 경우  <math display="block">\frac{(n+1)}{2} \text{번째 값}</math> </li> <li>■ 데이터의 개수가 짝수인 경우  <math display="block">\frac{n}{2} \text{번째 값과 } \frac{n}{2} + 1 \text{번째 값의 산술평균}</math> </li> </ul>
최빈값	주어진 데이터 중에서 가장 많이 나오는 값

● 산포도: 데이터의 흩어진 정도를 설명하는 통계치

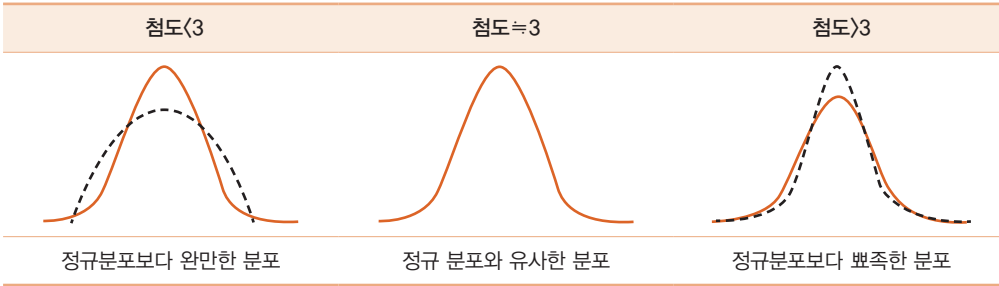
범위	데이터의 최대값에서 최소값을 뺀 것 $Range = Max - Min$
사분위수 범위	75% 지점에 위치한 값을 Q3(제3사분위수)에서 25% 지점에 위치한 값을 Q1(제1사분위수)을 뺀 값 $IQR = Q_3 - Q_1$
분산	편차를 제곱을 총합하여 평균 낸 값 $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
표준편차	분산 값에 제곱근을 씌운 값 $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$

● 왜도: 데이터 분포의 비대칭성을 나타내는 지표

positive skew (왜도>0)	negative skew (왜도<0)
<p>Histogram for capita crime rate by town</p>  <p>오른쪽으로 꼬리가 길다. (= 왼쪽에 데이터가 많은 형태)</p> <p>최빈값 &lt; 중앙값 &lt; 평균</p>	<p>Histogram for proportion of blacks by town</p>  <p>왼쪽으로 꼬리가 길다. (= 오른쪽에 데이터가 많은 형태)</p> <p>평균 &lt; 중앙값 &lt; 최빈값</p>



- **첨도:** 데이터들이 분포의 중심에 어느 정도 몰려 있는가를 측정할 때 사용하는 지표



## ■ 시각적 데이터 탐색 – 그래프 종류

히스토그램	연속형 변수 데이터의 도수분포를 보여주는 그래프
막대그래프	범주의 빈도를 직사각형 막대로 나타낸 그래프
줄기-잎 그림	데이터의 처음 몇 자릿수를 줄기로, 나머지는 잎으로 그린 표 형태와 그래프 형태가 혼합된 그래프
상자그림	다섯 숫자 요약을 그린 그래프로 아웃라이어 처리에 유용
산점도	직교좌표계에 서로 다른 연속형 변수의 값을 점을 찍은 그래프로, 두 연속형 변수의 관계 파악 시에 유용
원그래프	전체에 대한 각 부분의 비율을 원 모양으로 나타낸 그래프

## ■ R에서 자주 사용하는 시간 포맷 형식

형식	의미	예
%Y	4자리 연도	2020
%y	2자리 연도	20
%m	월 (01-12)	09
%b	월(영어축약형)	Jan
%B	월(영어명)	January
%d	일 (01-31)	12
%H	시(00-23)	16
%M	분(00-59)	19
%S	초(00-61)	32



## ■ 공간분석과 GIS

공간분석	사람들이 관심을 가지는 공간 데이터를 지도 위에 크기, 모양, 선의 굵기, 색상 등으로 구분해 시각화하여 인사이트를 얻는 분석 기법
지리정보시스템	지리 공간적으로 참조 가능한 모든 형태의 정보를 효율적으로 수집, 저장, 처리, 관리, 분석할 수 있게 설계된 컴퓨터의 하드웨어와 소프트웨어 및 지리적 자료, 인적 자원의 통합체로, 줄여서 GIS라고 함.

## ■ GIS 구성요소

- 컴퓨터 시스템
- GIS 소프트웨어
- 인력
- 데이터
- 인프라

## ■ 일변량 분석 vs. 이변량 분석 vs. 다변량 분석

일변량 분석	가장 간단한 형태의 분석으로, 1개의 변수를 대상으로 하며 데이터를 요약하거나 패턴을 찾는 것을 목표로 함. 일변량 데이터에서 패턴을 발견하기 위해 평균, 중앙값, 최빈값, 분산 등을 조사
이변량 분석	2개의 변수를 이용한 분석을 수행하며 두 변수 간의 관계를 주로 분석
다변량 분석	3개 이상의 변수를 이용한 복잡한 형태의 분석으로, 차원을 축소하거나 유사성 및 근접성을 기준으로 분류하는 식의 분석을 주로 수행



## ■ 다변량 분석 기법

상관분석	산점도 행렬을 그려 교차하는 변수 간의 관계를 보여주는 산점도와 상관계수를 파악
다차원 척도법	<p>객체 사이의 유사성 수준을 2차원 또는 3차원 공간에 점으로 시각화하는 분석 기법으로 거리 계산에 유클리드 거리를 주로 사용</p> $d_{xy} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$ <ul style="list-style-type: none"> <li>계량적 MDS: 거리를 계산하기 위해 유클리드 거리를 주로 활용하며 크게 데이터 간의 실제 거리를 근접도로 이용하는 전통적인 다차원 척도법</li> <li>비계량적 MDS: 순서 정보를 근접도로 이용하는 다차원 척도법</li> </ul>
주성분 분석	데이터의 분포를 잘 설명함과 동시에 정보의 손실은 최소화하게 고차원의 데이터를 저차원의 데이터로 변환하는 차원축소 기법
선형판별분석	어떤 그룹에 속할지를 판별하는 판별분석 기법으로 다변량 데이터에 판별 함수를 적용하여 데이터의 클래스 분리를 최적으로 수행할 수 있게 데이터를 축소

## ■ 텍스트 마이닝

텍스트 마이닝	다양한 문서 자료 내 비정형 텍스트 데이터에 자연어 처리 기술 및 문서 처리 기술을 활용해 인사이트를 도출하는 기술
자연어 처리	인간이 사용하는 언어(자연어)를 컴퓨터가 처리하고 분석할 수 있게 하는 작업으로 기계번역, 텍스트 분류, 질의응답 시스템, 스팸메일 검출 등의 영역에서 사용

## ■ 텍스트 마이닝 응용 분야

- 문서 요약: 문서 내 주제 및 핵심 내용을 추출하는 것
- 문서 분류: 문서의 내용에 따라 특정 카테고리 분류하는 것
- 문서 군집화: 성격이 유사한 문서들을 같은 군집으로 모아주는 것
- 특성 추출: 사용자가 원하는 의미 있는 특성을 추출하는 것

## ■ 텍스트 마이닝 용어

코퍼스	분석 작업의 대상이 되는 대량의 텍스트 문서를 모아놓은 집합
토큰화	구조화되어 있지 않은 문서를 단어(토큰)로 나누는 과정
불용어	코퍼스에서 자주 등장하지만 분석 프로세스에 있어 기여하는 바가 없는 단어로, 'is', 'a', 'the' 등이 이에 해당
어간 추출	단어 내 접사를 제거하고 단어에서 의미를 담고 있는 어간을 분리하는 것
표제어 추출	원형 단어를 찾는다는 점에서 어간 추출과 유사하지만, 그 단어가 어떤 품사로 쓰였는지가 지 고려한다는 점에서 차이점 존재
품사 태깅	문서 내 각 단어에 해당하는 품사로 태그를 달아주는 과정
형태소 분석	단위 형태소를 분리한 후에 변형이 일어난 형태소의 원형을 복원하고, 분리된 단위 형태소들로부터 단어 형성 규칙에 맞는 연속된 형태소들을 구하는 과정
N-그램	연속된 n개의 단어 혹은 형태소 집합
단어문서행렬	문서별로 나타난 단어의 빈도를 행렬 형태로 나타낸 것
TF-IDF	특징 추출의 기법으로 한 문서 내 특정 단어의 빈도를 나타내는 단어 빈도(term frequency, tf)와 전체 문서에서 단어가 몇 개의 문서에서 등장했는지를 나타내는 문서 빈도(document frequency)의 역수를 곱한 값
워드 클라우드	특정 문서에 사용된 단어로 구성된 구름 이미지로 각 단어의 크기는 출현 빈도와 중요성을 효과적으로 표현
토픽 모델링	대량의 문서 집합에 존재하는 추상적인 토픽(주제)을 추출하는 통계적 모델링 방법
잠재 디리클레 할당	가장 대표적으로 사용되고 있는 토픽 모델링 기법으로 데이터의 차원을 축소하는데 용이하고 의미가 일관된 토픽을 생성
Bag-of-Word	단어의 순서는 토픽과 상관이 없으며 단어의 빈도만이 중요하다는 개념

## ■ 소셜 네트워크 분석

소셜 네트워크 분석	소셜 네트워크 서비스 내 개인과 집단 간의 관계 및 상호작용을 모델링해 그것의 위상구조와 특성을 계량적으로 분석하고 시각화하는 방법론
------------	--



## ■ 소셜 네트워크 분석 방법론

집합론적 방법	객체와 관계를 집합 관계쌍으로 표현
그래프 방법	노드와 링크로 표현
행렬 방법	행렬의 행과 열에 객체를 대칭적으로 배치하면서 행렬의 $(i, j)$ 번째 위치에 $i$ 번째 객체와 $j$ 번째 객체 사이의 관계가 있고 없음을 1, 0으로 표현

## ■ 네트워크 구조를 파악하기 위한 요소 - 중심성

전체 네트워크에서 한 개체가 중심에 위치하는 정도를 표현하는 지표

연결 정도 중심성	한 노드에 직접 연결된 다른 노드들의 수의 합으로 중심성을 측정
근접 중심성	각 노드 간의 거리를 근거로 중심성을 측정
매개 중심성	네트워크 내에서 중계자 역할의 정도로 중심성을 측정
위세 중심성	연결된 노드의 중요성에 가중치를 두어 노드의 중심성을 측정

## ■ 네트워크 노드

밀도	네트워크 내에 존재하는 노드 간의 연결 정도의 수준
집중도	네트워크 전체가 한 중심에 집중되는 정도
연결 정도	노드에 연결된 관계의 수
포괄성	한 네트워크 내 연결되지 않은 노드들의 수를 뺀 연결된 노드들의 비율로, 포괄성이 높을수록 노드 간 관계가 많다고 해석



## ■ 기술 통계와 추론 통계

기술 통계	추론 통계
수집한 데이터를 요약, 묘사, 설명하는 통계 기법	수집한 데이터를 바탕으로 모수에 대하여 추론 또는 예측하는 통계 기법

## ■ 확률 표본 추출

단순 무작위 표본 추출	표본을 균등한 확률로 추출
체계 표본 추출	시간, 순서, 공간의 동일한 구간에서 무작위로 하나의 단위 추출, k번째 간격마다 추출
층화 표본 추출	여러 개의 층으로 나눈 후 각 층에서 표본을 단순 무작위로 추출
군집 표본 추출	하나의 군집 추출하여 일부 또는 전체를 조사

## ■ 비확률 표본 추출

편의 표본 추출	조사자가 표본 선정의 편리성에 기준을 두고 추출
판단 표본 추출	조사자의 주관적 판단으로 필요 대상만 추출
누적 표본 추출	사전에 알고 있는 대상을 조사하여 누적 표본 추출
할당 표본 추출	그룹화하여 그룹별로 필요한 대상만 추출

## ■ 확률분포: 확률변수의 분포 형태를 그래프로 표현한 것

확률 질량 함수	이산확률분포의 확률분포를 나타낸 함수
누적분포 함수	시작점을 음의 무한대로 통일한 특수 구간을 사용하는 함수 $P((-\infty \leq X \leq -1))$
확률 밀도 함수	임의의 지점의 밀도를 함수로 나타낸 것 (히스토그램 면적)



## ■ 이산확률분포

이항 분포	여러 번의 베르누이 시행 후 성공한 횟수를 확률변수로 하는 확률분포
다항 분포	각각의 경우가 나올 수 있는 횟수 집합의 분포
초기하 분포	이항분포는 복원 추출, 초기하 분포는 비복원 추출
포아송 분포	확률변수가 이항분포 $B(n, p)$ 를 따를 때 $np = \lambda$ 로 일정하게 두고, $n$ 이 충분히 크고 $p$ 가 0에 가까울 때 이항분포에 근사하는 분포

- 베르누이 시행: 연속된  $n$ 번의 독립적 시행에서 각 시행이 확률  $p$ 를 가질 때의 이산확률 분포

## ■ 연속확률분포

정규 분포	가우스 분포, 평균값 중앙으로 좌우 대칭인 종 모양의 분포
감마 분포	특정 수( $n$ )의 사건이 발생할 때까지 시간에 관한 분포
베타 분포	베타함수를 이용한 분포
t 분포	$t=0$ 에 대해 좌우 대칭을 이루는 종 모양의 분포, 자유도가 클수록 표준 정규 분포에 근사
카이 제곱 분포	정규 분포를 제곱 혹은 제곱한 것을 더해 나타낸 분포, 자유도가 커질수록 종 모양의 정규 분포에 근사
F 분포	두 데이터셋의 분산에 대한 분포

## ■ 표본 분포

표본 분포 개념	무수히 많은 표본의 평균값 혹은 표준편차에 대한 분포
중심 극한 정리	<ul style="list-style-type: none"> <li>■ 불규칙한 형태의 모집단에서 표본 크기가 <math>n</math>인 표본을 여러 번 반복 추출하여 정규 분포의 형태를 이루는 그래프로 변환하는 것을 의미</li> <li>■ 표본 수가 작아도 모집단 통계량을 추정할 수 있는 방법</li> </ul>

## ■ 점추정

추정	통계량을 통해 모집단의 모수를 추측하는 것
점추정	모집단의 모수를 단일 값으로 추정하는 것
추정량	모수를 추정하는 통계량
추정치	계산된 추정량의 값

## ■ 좋은 추정량의 조건

불편성	추정량의 기댓값이 모수의 실제값과 같거나 가까울수록 더 좋은 추정량
효율성	모든 불편 추정량 중에서 분산이 작을수록 더 좋은 추정량
일치성	표본의 크기를 크게 할수록 추정량은 모수에 가까워짐
충분성	모수에 대한 정보를 많이 제공할수록 더 좋은 추정량

## ■ 구간추정

구간추정	모수가 포함될 것으로 기대되는 구간을 추정하는 것이며, 신뢰성 정도를 포함
신뢰구간	구간추정을 통해 얻은 구간 $[\hat{\theta}_L, \hat{\theta}_U]$ 으로 $\hat{\theta}_L$ 은 신뢰 하한(lower confidence limit), $\hat{\theta}_U$ 은 신뢰 상한(upper confidence limit)을 의미
신뢰수준	n번 표본을 추출해서 구한 n개의 신뢰구간 중 모수를 포함하는 신뢰구간의 비율
오차율	오차율 $\alpha$ 는 신뢰구간에 모수가 포함되지 않을 확률

## ■ 신뢰구간과 z 값

- 90% 신뢰구간은  $\alpha$ 가 0.1인 경우로  $z_{\alpha/2} = z_{0.05} = 1.645$
- 95% 신뢰구간은  $\alpha$ 가 0.05인 경우로  $z_{\alpha/2} = z_{0.025} = 1.96$
- 99% 신뢰구간은  $\alpha$ 가 0.01인 경우로  $z_{\alpha/2} = z_{0.005} = 2.576$

## ■ 구간추정량 계산 방법

모분산 $\sigma^2$ 이 알려져 있는 경우	$\left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$
모분산 $\sigma^2$ 을 모르는 경우: 대표본	$\left( \bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right)$
모분산 $\sigma^2$ 을 모르는 경우: 소표본	$\left( \bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}} \right)$



표준오차를 아는 경우

$$(\bar{X} - z_{\alpha/2} SE, \bar{X} + z_{\alpha/2} SE)$$

## ■ 모비율의 신뢰구간 추정

$$\left( \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

## ■ 구간추정 성질

- ① 신뢰수준은 높으면서 신뢰구간은 좁은 것이 바람직
- ② 신뢰수준을 높이면 신뢰구간의 크기가 넓어져 정밀한 구간추정이 어려움
- ③ n이 클수록, 즉 표본의 크기가 충분히 클수록 신뢰구간이 좁아짐

## ■ 가설검정 용어

통계적 가설검정	모집단의 특성에 대한 주장 또는 가설을 세우고 표본에서 얻은 정보를 이용해 가설이 옳은지를 판정하는 과정
귀무가설	실험, 연구를 통해 기각시키고자 하는 어떤 가설에 해당하며 $H_0$ 로 표시
대립가설	실험, 연구를 통해 증명하고자 하는 새로운 아이디어 혹은 가설로 귀무가설을 기각함으로써 대립가설을 채택
검정통계량	가설의 검정에 사용되는 표본 통계량으로 결론을 내릴 때 사용되는 판단 기준
유의수준	귀무가설이 참인데도 이를 잘못 기각하는 오류를 범할 확률의 최대 허용 한계로 1%(0.01), 5%(0.05)를 주로 사용
기각역	귀무가설을 기각하게 될 검정통계량의 영역
채택역	귀무가설을 기각할 수 없는 검정통계량의 영역
임계값	기각역의 경계값
유의확률	귀무가설을 지지하는 정도를 나타낸 확률로 p-value라고도 함



## ■ 단일 검정과 양측 검정

대립가설은 단측검정과 양측검정으로 구분할 수 있다. 단측검정(one-sided test)은 검정통계량 분포의 왼쪽 끝 혹은 오른쪽 끝에 기각역이 존재하며 양측검정(two-sided test)은 검정통계량 분포의 양쪽 끝에 기각역이 존재하는 검정이다.

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0 \rightarrow \text{양측 검정(two-tailed)}$$

$$H_1: \mu < \mu_0 \rightarrow \text{단측 검정(left-tailed)}$$

$$H_1: \mu > \mu_0 \rightarrow \text{단측 검정(right-tailed)}$$

## ■ 가설검정과 오류

	$H_0$ 가 참	$H_0$ 가 거짓
$H_0$ 기각	제1종 오류( $\alpha$ )	옳은 결정( $1-\beta$ )
$H_0$ 채택	옳은 결정( $1-\alpha$ )	제2종 오류( $\beta$ )

## ■ 제1종 오류와 제2종 오류

### ① 제1종 오류

귀무가설이 참일 때 귀무가설을 기각하는 오류

### ② 제2종 오류

귀무가설이 거짓일 때 귀무가설을 채택하는 오류

## ■ 단일 모평균 검정

㉠ 모분산  $\sigma^2$ 를 아는 경우

대립가설	검정통계량	기각역
$H_1: \mu \neq \mu_0$		$z < -z_{\alpha/2}$ 또는 $z > z_{\alpha/2}$
좌측검정 $H_1: \mu < \mu_0$	$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$	$z < -z_\alpha$
우측검정 $H_1: \mu > \mu_0$		$z > z_\alpha$



② 모분산  $\sigma^2$ 를 모르는 경우

대립가설	검정통계량	기각역
$H_1: \mu \neq \mu_0$		$t < -t_{n-1, \alpha} \text{ 또는 } t > t_{n-1, \alpha/2}$
좌측검정 $H_1: \mu < \mu_0$	$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}, v = n - 1$	$t < -t_\alpha$
우측검정 $H_1: \mu > \mu_0$		$t > t_\alpha$

■ 단일 모비율 검정

대립가설	검정통계량	기각역
$H_1: p \neq p_0$		$z < -z_{\alpha/2} \text{ 또는 } z > z_{\alpha/2}$
좌측검정 $H_1: p < p_0$	$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$z < -z_\alpha$
우측검정 $H_1: p > p_0$		$z > z_\alpha$

■ t 검정

단일표본 t 검정	하나의 모집단의 평균값을 특정 값과 비교하는 경우 사용하는 통계적 분석 방법
독립표본 t 검정	서로 독립적인 두 그룹의 평균 차이가 0인지 알아보는 검정 방법
대응표본 t 검정	동일한 대상에 대해 두 가지 관측치가 있는 경우 이를 비교하여 차이가 있는지 검정할 때 사용