

☞ Descriptive Statistics & Inferential Statistics (1)

- 통계(statistics)의 목적에 따른 분류

- ✓ 기술 통계(Descriptive Statistics)
- ✓ 추리(추론) 통계(Inferential Statistics)

- 기술 통계(Descriptive Statistics)

- ✓ 수집한 데이터를 요약, 묘사, 설명하는 통계 기법
- ✓ 기술 통계는 다시 2가지 형태로 구분
- ✓ 집중화 경향(Central tendency)에 대한 기법 : 데이터를 대표하는 값이 무엇인지 혹은 어떤 값에 집중되어 있는지를 다루는 기법 (평균-mean, 중앙값-median, 최빈값-mode 등)
- ✓ 분산도(Variation)에 대한 기법 : 데이터가 어떻게 분포하고 있는지 설명하는 기법 (표준편차, 사분위)

☞ Descriptive Statistics & Inferential Statistics (2)

● 기술 통계 예시

- ✓ 기술 통계기법을 이용하면 수집한 데이터의 전체적인 형태를 알 수 있다.

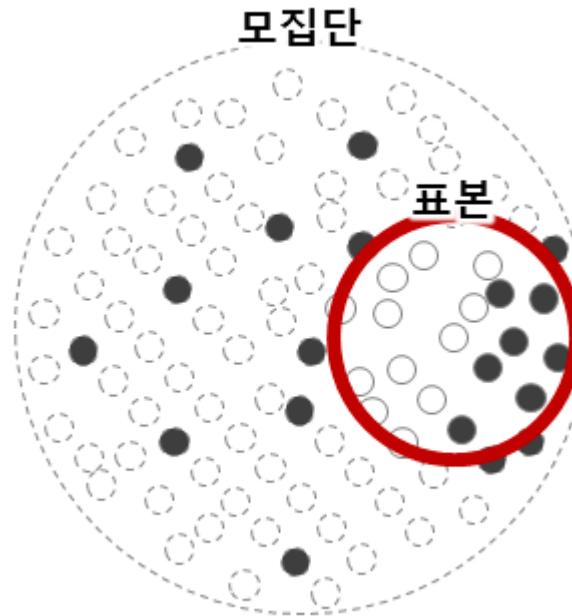
- ✓ 우리나라 국민 소득 데이터를 이용하면,
- ✓ 국민 1인당 평균 소득을 알 수 있고 (우리나라 국민 소득 수준의 대표값으로 활용)
- ✓ 국민 1인당 평균 소득에 대한 편차도 알 수 있고 소득의 편차가 크면 소득의 분포가 넓게 퍼져 있다는 의미
- ✓ 따라서 “소득의 분배가 잘 이루어지고 있지 않음” 을 알 수 있다.

☞ Descriptive Statistics & Inferential Statistics (3)

- 추리(추론) 통계(Inferential Statistics)

- ✓ 수집한 데이터를 기반으로 어떠한 사실을 예측(추론)하고 검정하는데 사용하는 통계 기법

- ✓ 대통령 선거 예측



☞ Descriptive Statistics & Inferential Statistics (4)

● 추리(추론) 통계(Inferential Statistics)

- ✓ 현재는 빅데이터의 개념과 함께 모집단과 표본집단을 약간 다르게 해석.
- ✓ 내가 가지고 있는 데이터 전체를 표본집단, 아직 수집하지 못한 데이터 혹은 미래에 발생하는 데이터를 모집단으로 간주.
- ✓ 현재 우리 회사 제품 구매 고객 데이터 => 표본집단
- 타 회사 제품 구매 고객 데이터 & 미래에 제품을 구매할 고객 데이터 => 모집단

☞ Descriptive Statistics & Inferential Statistics (5)

● 추리(추론) 통계(Inferential Statistics)

- ✓ 추리 통계에서는 다음과 같은 3가지 사항에 항상 주의 해야 한다.
- ✓ 표본집단은 모집단을 대표할 수 있는가?
- ✓ 표본의 확률분포는 무엇인가? => 표본의 수가 많아질수록 정규분포에 접근하게 된다.
- ✓ 추정된 결과는 신뢰성이 있는가?

☞ Gaussian normal distribution (참고)

- Gaussian normal distribution (가우시안 정규분포) : 일반적으로 간단히 정규분포라고 한다.

- ✓ 자연현상에서 나타나는 숫자를 확률 모형으로 모형화 할 때 사용
- ✓ 정규분포는 평균(μ)과 분산(σ^2)이라는 두 개의 모수만으로 정의되며 확률밀도함수(PDF – probability density function)는 다음과 같은 수식으로 표현

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

- ✓ 정규분포 중에서 평균이 0이고 표준편차가 1인 정규분포를 표준정규분포(standard normal distribution)이라고 한다.

☞ 통계적 가설 검정 (Statistical Hypothesis Testing)

- 표본에서 얻은 사실을 근거로 모집단에 대한 가설이 맞는지 통계적으로 검정하는 분석방법
 - ✓ 증명되지 않은 주장이나 가설을 표본 통계량에 입각하여 주장이나 가설의 진위 여부를 판단, 증명, 검정하는 통계적 추론 방식
- 가장 먼저 해야 할 일은 가설(Hypothesis)을 설정하는 것
 - ✓ 귀무가설(null hypothesis) : 진실일 가능성이 적어 reject 될 것이 예상되는 가설로 기각(reject)이 목표인 가설 (일반적으로 H_0 로 표현) => 차이가 없다. 영향력이 없다. 연관성이 없다. 효과가 없다.
 - ✓ 대립가설(alternative hypothesis) : 귀무가설에 대립되는 가설로 새로운 주장 혹은 실제로 입증하고 싶은 가설. 귀무가설이 기각될 때 받아들여지는 가설로 채택(accept)이 목표인 가설 (일반적으로 H_1 로 표현) => 차이가 있다. 영향력이 있다. 연관성이 있다. 효과가 있다.

☞ 통계적 가설 검정 (Statistical Hypothesis Testing)

- 어떤 회사 A의 지난 해 평균 월급은 150만원이고, 표준편차는 12만원이다. 올해는 임금이 인상되어 그보다 높을 것이라 예상되어 임의의 직원 100명을 뽑아 평균을 조사했더니 152만원 이었다. 이 때, 올해 평균 월급이 150만원보다 높다고 할 수 있는가? (우측검정)

- ✓ 귀무가설 (H_0) : $\mu = 150$ 만원
- ✓ 대립가설 (H_1) : $\mu > 150$ 만원



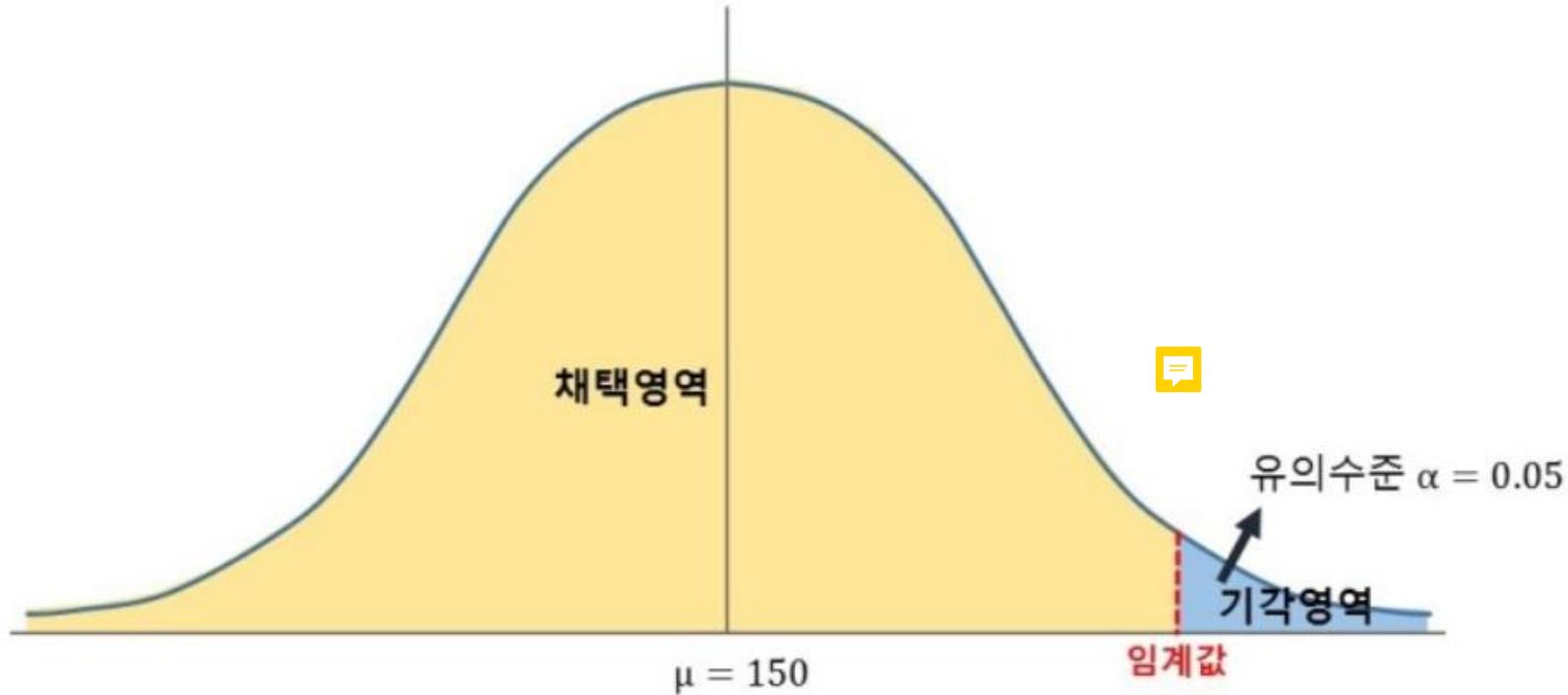
☞ 통계적 가설 검정 (Statistical Hypothesis Testing)



- 임계값(critical value) : 주어진 유의수준에서 귀무가설이 채택되는지, 기각되는지에 대한 기준값.
- 유의 수준(level of significance) : 귀무가설이 실제로 맞음에도 기각할 확률. 일반적으로 주어진 값을 이용한다. ($\alpha = 0.05$)
- 어떤 회사 A의 지난 해 평균 월급은 150만원이고, 표준편차는 12만원이다. 올해는 임금이 인상되어 그보다 높을 것이라 예상되어 임의의 직원 100명을 뽑아 평균을 조사했더니 152만원 이었다. 이 때, 올해 평균 월급이 150만원보다 높다고 할 수 있는지를 유의수준 $\alpha = 0.05$ 에서 결정하세요.

☞ 통계적 가설 검정 (Statistical Hypothesis Testing)

- 그래프로 표현



☞ 통계적 가설 검정 (Statistical Hypothesis Testing)



- 귀무가설이 accept인지 reject인지를 판단하기 위한 기준인 임계값을 계산
 - ✓ 유의수준 $\alpha = 0.05$ 이므로 구하고자 하는 확률영역은 0.95.
 - ✓ 표준정규분포표를 이용해 95%에 해당하는 z값을 찾는다. ($z = 1.645$)

$$Z = \frac{CV - mean}{\sqrt{\frac{standard deviation}{number of sample}}}$$

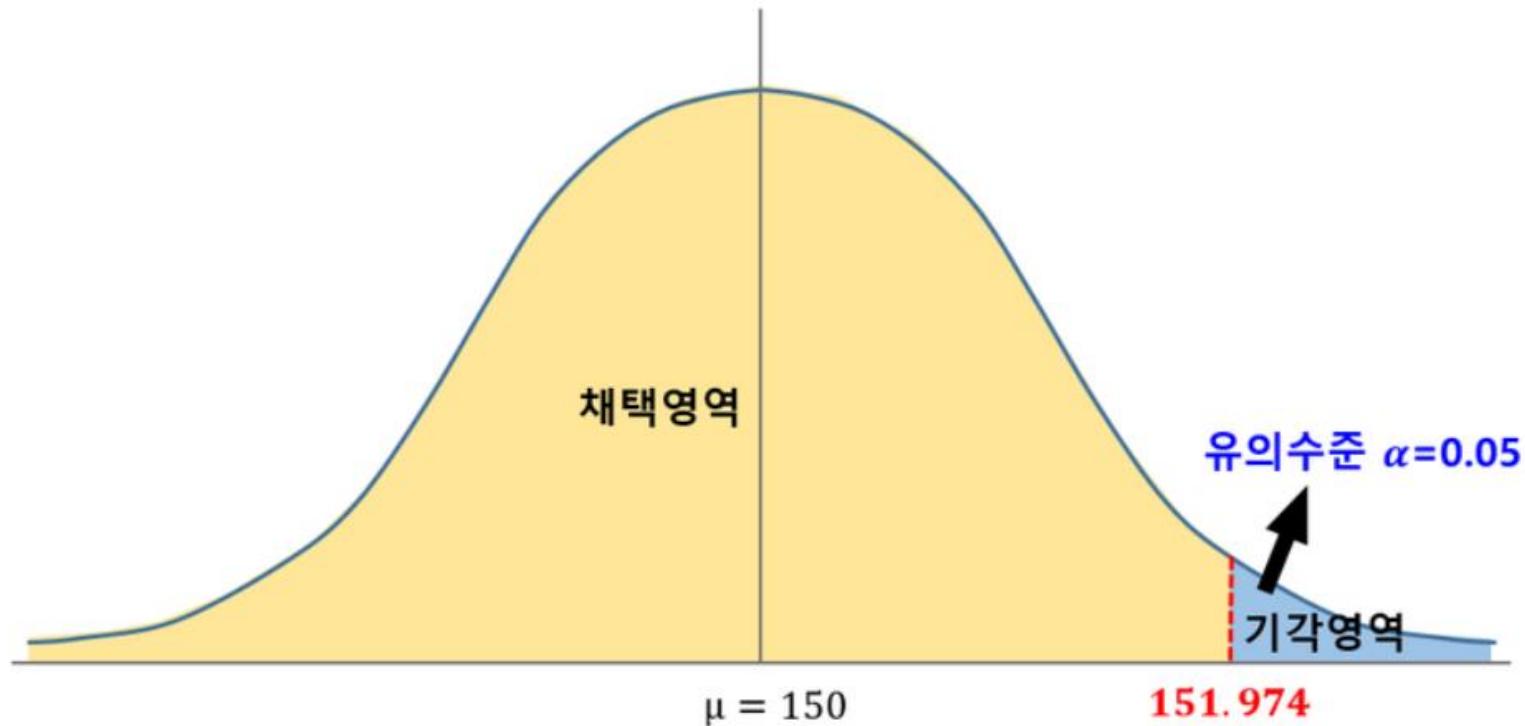
$$1.645 = \frac{CV - 150}{\sqrt{\frac{12}{100}}}$$

α	Z값
0.1	1.28
0.05	1.64
0.025	1.96
0.01	2.33
0.005	2.58

$$CV = 151.974$$

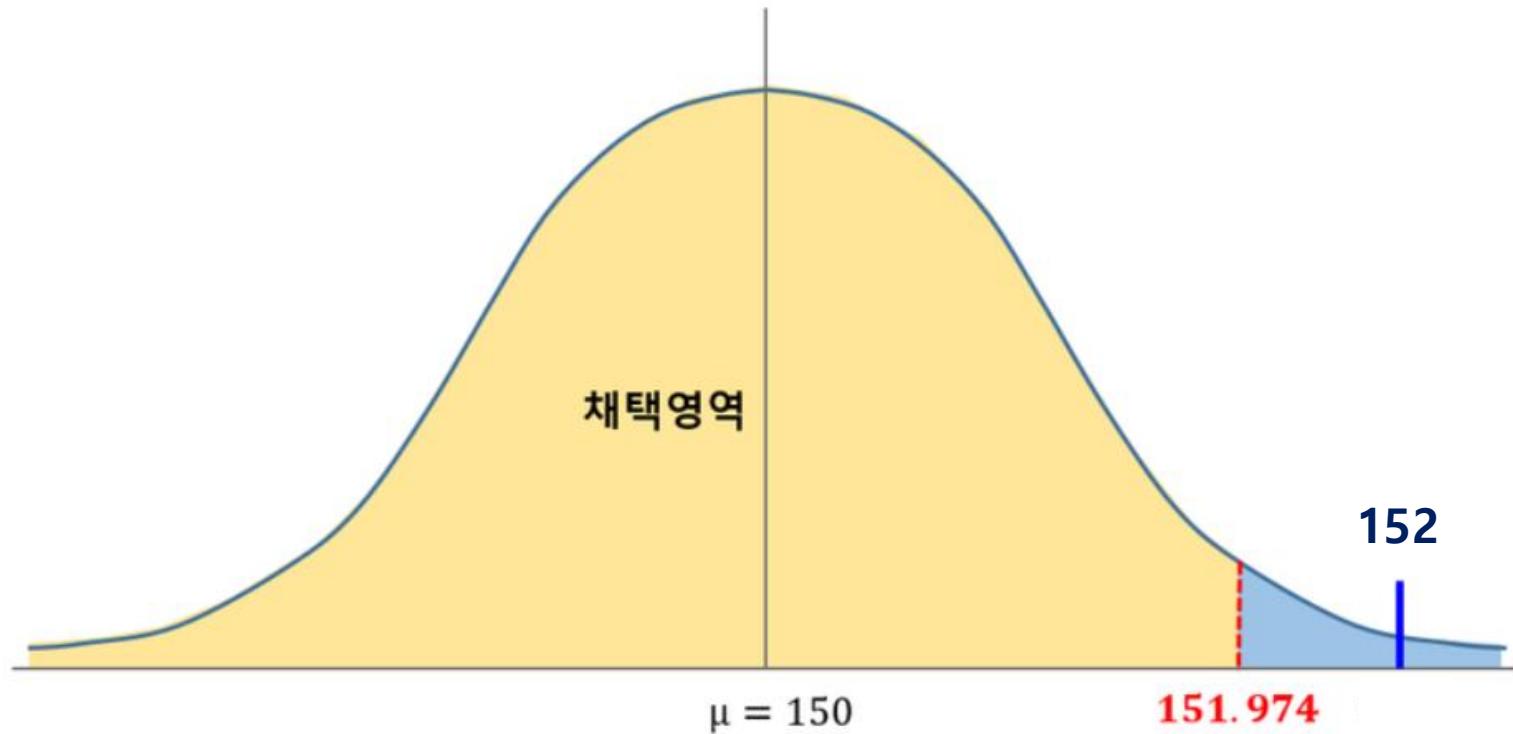
☞ 통계적 가설 검정 (Statistical Hypothesis Testing)

- 표본 직원 100명의 월급 평균은 152만원. 임계값은 151.974
 - ✓ 표본평균 152만원은 reject 영역에 포함



☞ 통계적 가설 검정 (Statistical Hypothesis Testing)

- 표본 평균이 임계값보다 크기 때문에 귀무가설이 기각
 - ✓ 자동으로 대립가설이 채택



☞ 통계적 가설 검정 (Statistical Hypothesis Testing)

☞ 통계적 가설 검정 (Statistical Hypothesis Testing)

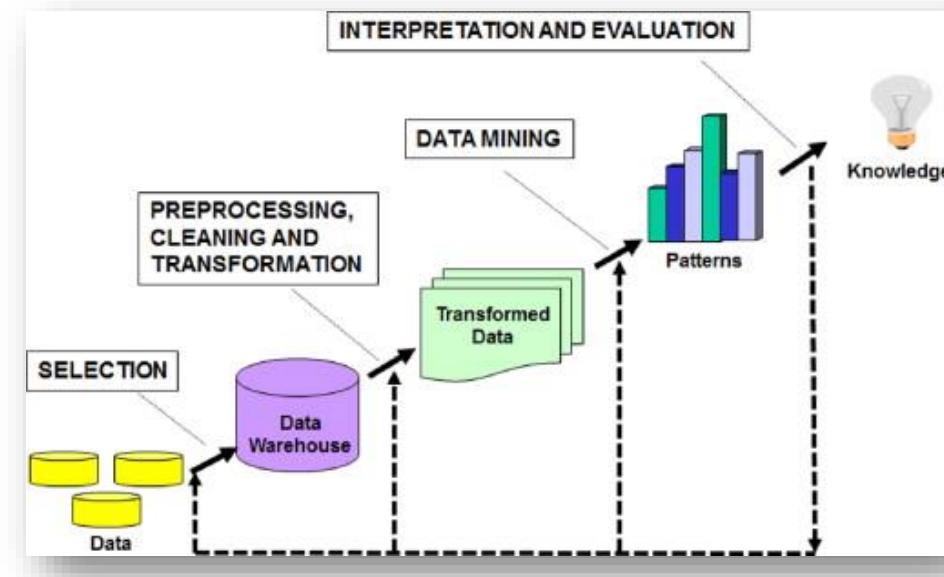
- 우리나라 남자들의 평균 수명은 75라고 한다. 하지만 의학기술의 발달로 평균 수명이 더 높아졌을 것이라는 의견이 나오고 있다. 그래서 최근에 사망한 남성 30명의 평균 수명을 조사하였더니 79세로 나왔다고 한다. 이때 어느 가설이 더 타당한지를 유의수준 10%에서 검정하세요. 단, 모집단의 표준편차는 10이다.

- ✓ 열심히 계산해보면 남성의 평균 수명은 75세보다 크다는 대립가설이 채택됨을 알 수 있다.
- ✓ 프로그램으로 구현

☞ 통계적 가설 검정 (Statistical Hypothesis Testing)

☞ Data Mining

- 대용량의 데이터로부터 유용한 정보를 추출하는 작업
- 대용량 데이터에 존재하는 데이터 간의 관계, 패턴, 규칙을 찾아내고 모형화 해서 의사 결정을 돋는 일련의 과정
- KDD (Knowledge Data Discovery)
Data로부터 유용한 지식(knowledge)을 찾는 과정



☞ Machine Learning

- Machine Learning은 Software.
- Explicit programming의 한계때문에 주목을 받음.
 - Explicit Programming은 우리가 일반적으로 하는 프로그래밍
 - Explicit Programming으로 해결할 수 없는 문제들이 있음. (경우의 수가 너무 많음)
 - 대표적인 것들로 Email Spam Filter, 자율 주행 시스템, 바둑 프로그래밍 등.
- 1959년 Arthur Samuel에 의해서 시작.
 - Machine Learning : 프로그램 자체가 데이터를 기반으로 학습을 통해 배우는 능력을 가지는 프로그래밍

☞ Data Mining vs. Machine Learning

- Data Mining : 통계학적으로 다양한 관점에서 데이터를 분석하여 의미를 도출

- Data Mining 기법 : 연관, 회귀, 분류라는 3가지 유형
- 연관 (association) : Data set에서 자주 발생하는 속성값들을 연결해주는 연관규칙을 발견하는 작업
(마트의 고객 쇼핑 카트 내의 개별 상품 간의 상관관계를 식별)
- 회귀 (regression) : 독립변수 분석을 통해 종속변수를 예측하는 작업
(고객의 소득수준과 상품 가격의 상관관계를 이용해 상품의 예상 매출액을 예측)
- 분류 (classification) : 개체들을 여러 등급으로 나누는 모델.

- Machine Learning : 미래 사건의 결과를 예측하는 컴퓨터 프로그래밍 (CS관점)

- 지도학습 : regression
- 비지도학습 : clustering
- 강화학습

☞ Machine Learning / Data Mining의 기법 (1)

- 분류 모형 (Classification Models)
- 전처리 (Preprocessing Methods)
- 군집화 (Clustering Methods)
- 최적화 (Optimization Methods)
- 가치평가 (Valuation Methods)

☞ Machine Learning / Data Mining의 기법 (2)

● 분류 모형 (Classification Models)

- ✓ 특정 기준에 근거해 분석 대상을 2개 혹은 그 이상의 집단으로 분류하는 예측 Model 
- ✓ 예) 날씨예측, 주가예측, 기업부도예측

- ✓ 다중판별분석 (MDA : Multiple Discriminant Analysis)
- ✓ 로지스틱 회귀분석(LOGIT : Logistic Regression) 
- ✓ 인공신경망 (ANN : Artificial Neural Networks)
- ✓ 사례기반추론 (CBR : Case-Based Reasoning)
- ✓ 의사결정트리 (DT : Decision Tree)
- ✓ SVM (Support Vector Machine)

☞ Machine Learning / Data Mining의 기법 (3)

● 최적화 (Optimization Methods)

- ✓ 주어진 제약조건 하에서 특정 목적 함수를 최대, 최소화 하는 변수들의 최적 값을 도출하는 기법
- ✓ 예) 공장의 생산량을 최대화 문제, 유통비용 최소화 문제



- ✓ 선형 계획법 (LP : Linear Programming)
- ✓ 유전자 알고리즘 (GA : genetic Algorithms)

● 전처리 (Preprocessing Methods)

- ✓ 예측 Model의 성과를 향상시키기 위해 입력데이터에 대해 사전 처리를 수행하는 기법
- ✓ 주성분분석(PCA : Principal Component Analysis)
- ✓ 퍼지 이론 (Fuzzy theory)

☞ Machine Learning / Data Mining의 기법 (4)

● 군집화 (Clustering Methods)

- ✓ 사전에 정해진 기준없이 서로 유사한 데이터들을 같은 그룹으로 묶어주는 기법
- ✓ 예) 고객 세분화 (우량고객, 불량고객), News grouping (정치, 경제, 연예, 스포츠)
- ✓ k-means 분류기법 (k-means clustering)

● 가치평가 (Valuation Methods)

- ✓ 정성적 측정대상에 대한 가치를 비교, 평가하는 기법
- ✓ 예) 전자제품(핸드폰) 선정
- ✓ 분석적 계층 프로세스 (AHP : Analytic Hierarchy Process)

☞ Data Preprocessing (1)

● 결측값(Missing Value) 처리

- ✓ 항목이 비어 있는 field, 너무 많은 field의 값이 비어 있는 record는 사용하기 힘듬
- ✓ 삭제하는 방법, 평균값, 중앙값, 최빈치를 이용하여 결측값을 대체해서 사용

● 정성적 변수의 정량화

- ✓ 각 field의 값은 atomic value를 가지도록 처리해야 한다.
- ✓ 성별이나 주소와 같은 정성적 parameter는 해석이 불가능 하기 때문에 binary code 형태로 변환해서 사용

● 이상치(Outlier) 제거

- ✓ 상식적으로 맞지 않는 값이거나 잘못 입력된 것으로 추정되는 변수의 값을 조절
- ✓ 상위 10%, 하위 10%에 해당하는 값을 특정 값으로 조정
- ✓ 예) 나이가 70이상은 무조건 70으로 설정 (이상치 값들은 가중치를 많이 부여 받는 성향이 있기 때문)

☞ Data Preprocessing (2)

● 정규화 (Normalization)

- ✓ 모든 입력 parameter의 값이 0과 1사이의 값을 가지도록 조정하면서 데이터가 표준 정규분포를 가지도록 값을 조정
- ✓ 입력 parameter 값의 편차가 크면 학습이 잘 진행되지 않거나 결과가 왜곡될 확률이 많다.
- ✓ 일반적으로 Min-Max Scale을 많이 사용 => $(x - \text{최소값}) / (\text{최대값} - \text{최소값})$

● 과적합화(Overfitting)이 발생되지 않도록 Hold-out data(검증데이터)를 이용

- ✓ training data set과 test data set을 7:3 혹은 8:2로 사용
- ✓ binary 예측의 경우 0과 1의 비중이 각 data set마다 1:1 비율이 되도록 사용

☞ Data Preprocessing (3)

● 모형에 들어갈 입력변수 후보 선정

- ✓ 모든 변수들이 의미 있는 변수는 아니다. 변수들이 의미가 있느냐 없느냐를 판단해야 한다.
- ✓ 카이제곱 검정(Chi-squared Test)
- ✓ 독립표본 t검정 (2 sample t-test), 대응표본 t검정 (paired t-test)
- ✓ 분산분석 (ANOVA : Analysis of Variance)

☞ 카이제곱 검정 (Chi-squared Test)

- 범주별로 빈도만 구해진 범주형 데이터에 대해 시행하는 독립성 검정 방법
- 두 범주형 데이터에 대해 서로 연관성이 있는가 혹은 그렇지 않은가를 알아보기 위한 검증방법
- 예시) 흡연량과 음주량 사이에 연관성이 있는가? 

	1갑 이상	1갑 이하	안피움	계
반병 이상	23	21	63	107
반병 이하	31	48	159	238
못마심	13	23	119	155
계	67	92	341	500

☞ 카이제곱 검정 (Chi-squared Test)

- 귀무가설 : 흡연과 주량은 연관성이 없다. 
- 대립가설 : 흡연과 주량은 연관성이 있다. (독립적이 아니다. – 주장하고 싶은 내용)
- 범주별 기대값 구하기 (1갑 이상 & 반병 이상)
 - ✓ 전체대상 500명 중 1갑 이상 담배를 피는 사람의 수는 67명
 - ✓ 전체대상 500명 중 반병 이상 술을 마시는 사람수는 107명
 - ✓ 500명 중 담배 1갑 이상 술 반병 이상 마시는 경우의 확률

$$\frac{67}{500} \times \frac{107}{500} = 0.02867$$



- ✓ 기대값은 확률과 해당 사건의 총 수를 곱해서 계산되므로 기대값은 다음과 같다.

$$0.02786 \times 500 = 14.338$$

☞ 카이제곱 검정 (Chi-squared Test)

- 모든 범주에 대한 기대값을 계산

	1갑 이상	1갑 이하	안피움	계
반병 이상	23 14.338	21 19.688	63 72.974	107
반병 이하	31 31.892	48 43.792	159 162.316	238
못마심	13 20.77	23 28.52	119 105.71	155
계	67	92	341	500

☞ 카이제곱 검정 (Chi-squared Test)

- 각 범주별

카이제곱값을
구한다.

$$\chi^2 = \frac{(\text{관측값} - \text{기대값})^2}{\text{기대값}}$$

	1갑 이상	1갑 이하	안피움	계
반병 이상 기대값 관측값-기대값 제곱 카이스퀘어	23 14.338 8.66 75.03 5.23	21 19.688 1.31 1.72 0.09	63 72.974 -9.97 99.48 1.36	107
반병 이하 기대값 관측값-기대값 제곱 카이스퀘어	31 31.89 -0.89 0.80 0.02	48 43.792 4.21 17.71 0.40	159 162.316 -3.32 11.00 0.07	238
못마심 기대값 관측값-기대값 제곱 카이스퀘어	13 20.77 -7.8 60.4 2.9	23 28.52 -5.5 30.5 1.1	119 105.71 13.3 176.6 1.7	155
계	67	92	341	500

☞ 카이제곱 검정 (Chi-squared Test)

● 전체 카이제곱 값 합하기

- ✓ $(5.23+0.09+1.36+0.02+0.4+0.07+2.9+1.1+1.7) = 12.87$

● 카이제곱의 자유도 구하기

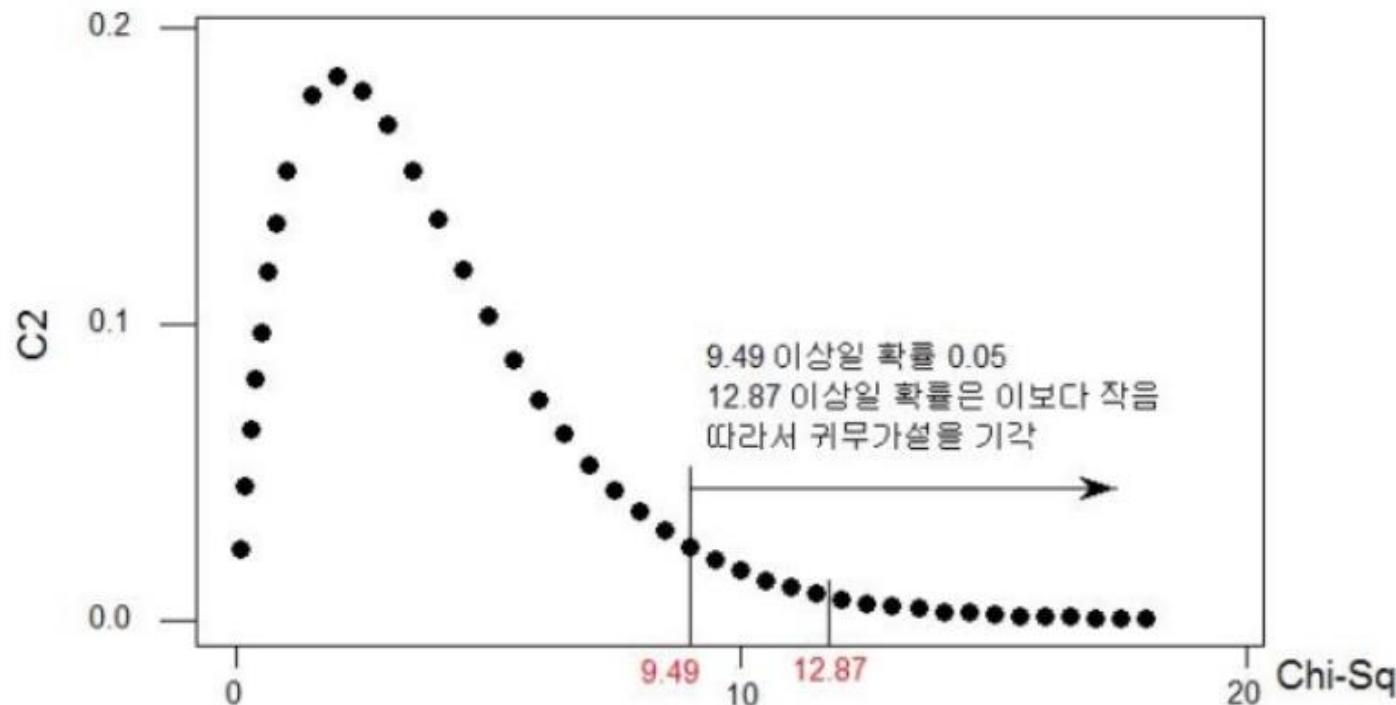
- ✓ 흡연의 자유도 : 2, 주량의 자유도 : 2
- ✓ 전체 자유도는 각각의 자유도의 곱 : $2 \times 2 = 4$

● 검정 결과 도출

- ✓ 카이제곱 분포표를 기준으로 자유도와 α 값을 이용하여 카이제곱 값을 알아온다.
- ✓ 자유도(4)와 α 값(0.05) $\Rightarrow 9.49$
- ✓ 우리가 구한 카이제곱 값은 12.87 \rightarrow 카이제곱 검정값은 9.49
- ✓ 따라서 귀무가설을 기각하고 대립가설을 채택

☞ 카이제곱 검정 (Chi-squared Test)

- 그래프로 표현하면 다음과 같다.



☞ 카이제곱 검정 (Chi-squared Test)

☞ 독립표본 t검정 (2 sample t-test)

- 두 집단간의 평균값의 차이가 통계적으로 의미가 있는지를 검정하는 방법 

- ✓ 두 집단간 평균값의 차이가 우연한 차이인지 아니면 어떠한 의미가 있는 차이인지를 검정

- 예시)

- ✓ A학교 학생들과 B학교 학생들의 수학 시험 평균 점수의 차이가 통계적으로 유의한가?
 - ✓ 동일한 제품을 생산하는 A공장의 일일 생산량과 B공장의 일일 생산량의 차이가 통계적으로 유의한가?
 - ✓ 우리나라 30대 남성과 여성의 평균 신장의 차이가 통계적으로 유의한가?

- 각 변수는 기본적으로 정규분포여야 하며 분산의 동일성에 따라 공식이 달라진다.

☞ 독립표본 t검정 (2 sample t-test)

- 예시)

☞ 대응표본 t검정 (pared t-test)

- 특정 집단을 대상으로 실험 전과 실험 후의 차이가 통계적으로 의미가 있는지를 검정하는 방법
 - ✓ 즉, 실험 효과를 과학적으로 입증하고자 할 때 사용
- 예시)
 - ✓ 간기능 강화제 섭취 전과 후의 평균 주량에 차이가 있는가?
 - ✓ 논리적 사고의 향상을 위해 “코딩교육” 을 실시한 후 교육 전과 교육 후의 논리적 사고의 정도에 차이가 있는가?
 - ✓ 회사 워크샵 전과 후의 애사심에 차이가 있는가?
- 각 변수는 기본적으로 정규분포여야 한다.

☞ 대응표본 t검정 (pared t-test)

- 예시)

☞ 분산분석 (ANOVA)

- 2개의 데이터를 비교할 때는 t-test 이용. 3개 이상의 데이터를 비교할 경우 ANOVA를 이용하여 검정한다.
- 예시)
 - ✓ 고등학생들의 급식 만족도는 학년별로 차이가 있는가?
 - ✓ 4가지 교육훈련 기법이 차이가 있는가?
- 각 변수는 기본적으로 독립적인 변수이고 분석대상 종속변수는 정규분포여야 한다.
- 종속변수 분포는 분산의 동질성이 확보(등분산성)되어야 한다.

☞ 분산분석 (ANOVA)

- 예시)

☞ Regression Analysis (회귀분석)

- 통계학에서 중요한 역할을 담당하는 자료분석 방법.
- 출발점 : “관찰된 자료들이 어떤 특정한 경향성을 가지고 있지 않을까?” 라는 의문
- 관찰된 자료의 변수들 사이에 나타나는 경향성 혹은 의존성을 수학적으로 판별하고자 하는 기법
- 이러한 경향성이나 의존성을 발견한다면 앞으로 발생할 일들에 대한 예측(prediction)이 가능.

☞ Regression Analysis (회귀분석)

- 예) 여름철의 기온과 코카콜라의 판매량

- 두 변수 사이에 어떤 관계가 있을 것이라 생각되고 이에 따른 여러가지 가설(regression model)을 만들 수 있다.
- 만약, 우리가 얻은 회귀모형이 타당하다면 “독립변수와 의존변수 간에 의존관계가 있다” 라고 말할 수 있다.
- 또한 독립변수의 값을 토대로 종속변수의 값을 예측할 수 도 있다.

- 회귀분석에서 독립변수가 하나일 경우 단순회귀분석(simple regression), 독립변수가 둘 이상인 경우 다중회귀분석(multiple regression) 이라고 한다.

☞ Simple Linear Regression Model (단순선형회귀모델)

- 독립변수와 종속변수의 경향을 알아보는 가장 기본적인 방법은 1차 함수 (Linear function) 형태로 표현하는 것.

- 1차 함수는 종속변수와 독립변수 이외에 2가지 요소가 더 포함
- 기울기(slop), y절편 (y-intercept)
- 표현되는 수식은 다음과 같다.

$$y = \beta_0 + \beta_1 x$$

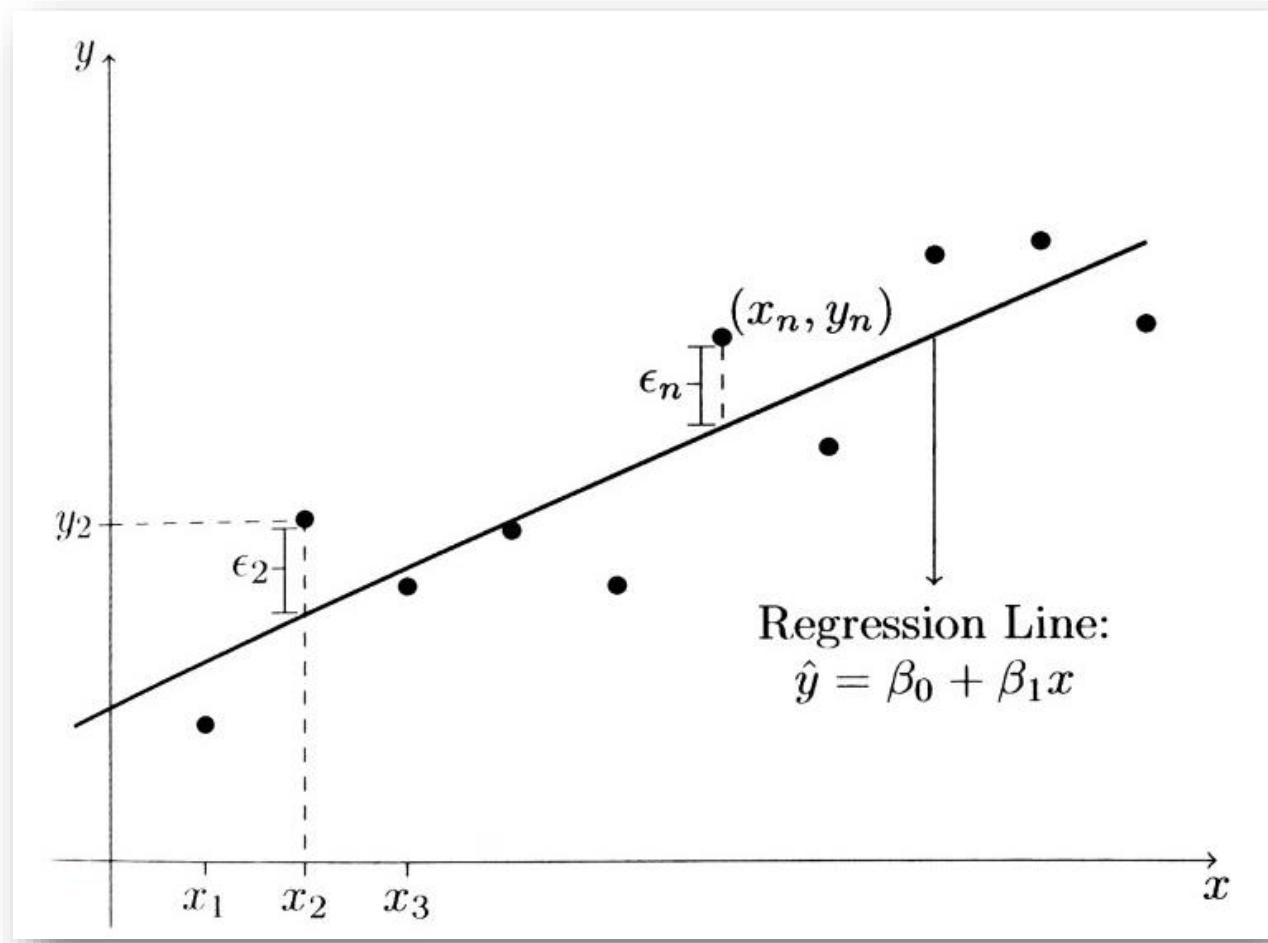
- 수학 공부한 시간과 수학시험성적의 관계

- 수학 공부한 시간을 독립변수 x로 설정
- 수학시험성적을 종속변수 y로 설정

☞ Simple Linear Regression Model (단순선형회귀모델)

- 사용할 데이터 : 학생 100명의 수학 공부시간과 수학시험성적
- 우리의 목표는 자료를 종합하여 오차가 가장 작은 일차함수를 만드는 것. 즉, 자료를 잘 표현하는 기울기와 절편을 구하는 것.
 - 이렇게 얻은 일차함수식을 회귀선(regression line) 혹은 단순선형회귀모델(simple linear regression model)이라고 한다.

☞ Simple Linear Regression Model (단순선형회귀모델)



☞ Simple Linear Regression Model (단순선형회귀모델)

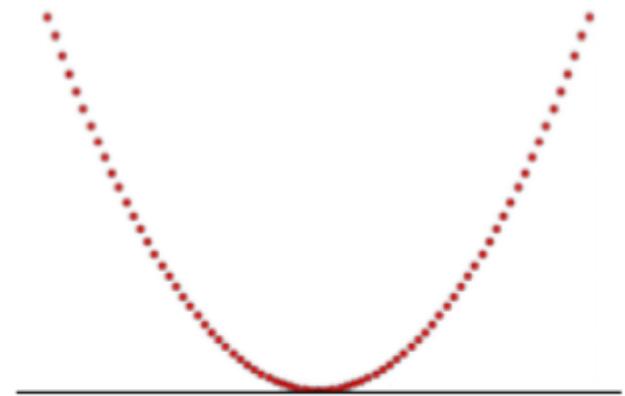
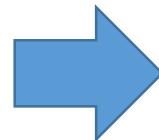
- 이전 그림과 같은 regression line을 만들려면 주어진 자료를 잘 표현하는 β_0 (y절편)와 β_1 (기울기)를 구해야 한다.
- β_0 (intercept coefficient)와 β_1 (slop coefficient)라고 불리는 regression coefficient(회귀계수)를 구하는 방법은 크게 2가지
 - 첫번째 방법은 기댓값과 공분산을 이용해 구하는 방법
 - 두번째 방법은 최소 제곱법
 - 여기서는 최소 제곱법만 이용해서 회귀계수를 구하는 개념에 대해서 설명한다.

☞ Simple Linear Regression Model (단순선형회귀모델)

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\frac{(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2 + (\hat{y}_4 - y_4)^2 + \dots + (\hat{y}_n - y_n)^2}{n}$$

$$\frac{\sum_{i=1}^n (\hat{y}_{(i)} - y_{(i)})^2}{n}$$



☞ Simple Linear Regression Model (단순선형회귀모델)

☞ Simple Linear Regression Model (단순선형회귀모델) – 연습문제

- 온도에 따른 ozone량 예측
- CSV 파일로 부터 온도와 온도에 따른 ozone량 데이터를 loading후 회귀분석을 통해 온도와 ozone량 간의 관계를 분석하고 온도에 따른 ozone량을 예측한다.
- 사용할 파일 : ozone.csv

☞ Simple Linear Regression Model (단순선형회귀모델) – 연습문제