방학 3주차

빅데이터를 위한 통계학

수나로움

빅데이터를 위한 통계학

통계의 종류

기술 통계 (Descriptive Statistics)	수집된 자료를 이용하여 표본의 통계량을 구하거나 자료를 요약하여 정보로 만드는 데 사용하는 통계
추론 통계 (Inferential Statistics)	표본의 통계량을 이용하여 모수를 추정하거나 모수에 대한 가설을 검정하는 데 사용하는 통계

빅데이터를 위한 통계학

통계 기초 개념 1

모수 (parameter) : 모집단 구성원 모두를 측정하여 얻을 수 있는 모집단의 특성

ㄴ 모평균(μ), 모분산(σ^2), 모표준편차(σ) 등

통계량 (statistic) : 표본의 관측치를 측정하여 얻은 값으로, 표본의 특성을 나타내는 값

ㄴ 표본평균(X bar), 표본분산(s^2), 표본표준편차(s) 등

통계치 : 특정한 표본의 통계량 값

빅데이터를 위한 통계학

통계 기초 개념 2

데이터 : 사람, 물건, 조건, 상황 등을 묘사하는 기본적인 사실의 집합

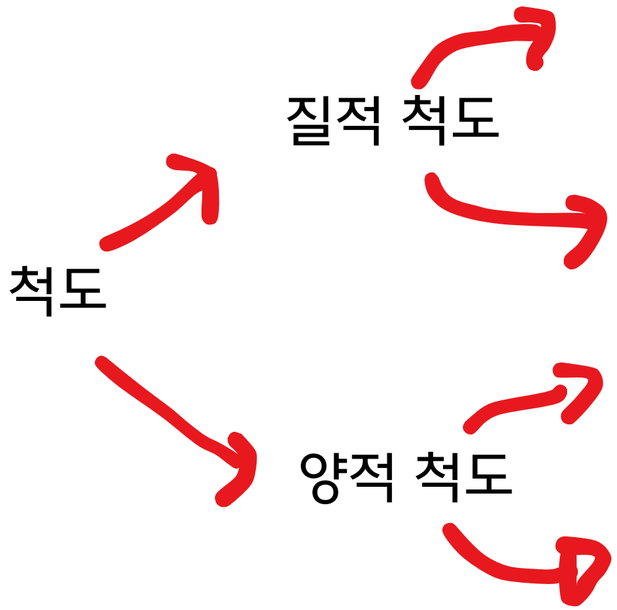
ㄴ 데이터(data)를 정보(information)으로 바꾸기 위해 노력함

자료 → 질적 자료 → 척도 → 측정 → 양적 자료

빅데이터를 위한 통계학

통계 기초 개념 2

척도 : 관찰 대상의 속성을 측정하여 그 값을 숫자로 나타내는 일종의 규칙



명목 척도 : 관찰대상을 단순히 범주로 분류하기 위한 척도

서열 척도 : 관찰대상의 크기를 바탕으로 순위를 나타낸 척도

등간 척도 : 상대적 크기를 측정해서 서로 비교할 수 있는 척도

비율 척도 : 절대적 기준인 비율을 통해 크기를 구하는 척도

빅데이터를 위한 통계학

통계 기초 개념 2

명목 척도 (nomial scale)

- 속성에 따라 관찰 대상을 상호배타적이고 포괄적인 범주로 구분하는 수치를 부여하는 도구
- 양적 의미 X => 그저 대상을 구분하는 도구

빅데이터를 위한 통계학

통계 기초 개념 2

서열 척도 (ordinal scale)

- 속성의 크기에 다라 관찰 대상의 순위를 나타내는 수치를 부여하는 측정 도구
- 상대적 순위만 구분하고, 그 안의 차이는 중요하지 X

2018~2020년 과학논문 순위 많이 인용된 순위

	상위 10%	10년전	상위 1%	10년전
1	중국	2 ↑	중국	3 ↑
2	미국	1 ↓	미국	1 ↓
3	영국	3 -	영국	2 ↓
4	독일	4 -	독일	4 -
5	이탈리아	8 ↑	호주	8 ↑
⋮				
10	스페인	10 -	일본	7 ↓
11	한국	13 ↑	스페인	11 -
12	일본	6 ↓	한국	14 ↑

빅데이터를 위한 통계학

통계 기초 개념 2

등간 척도 (interval scale)

- 관찰대상이 가진 속성의 순서 뿐만 아니라 상대적 차이도 고려하는 측정 도구
- 같은 간격으로 분할한 척도를 사용함

1. 다음은 논문 통계 공부에 있어 '스탯솔'의 도움 정도에 관한 질문입니다.

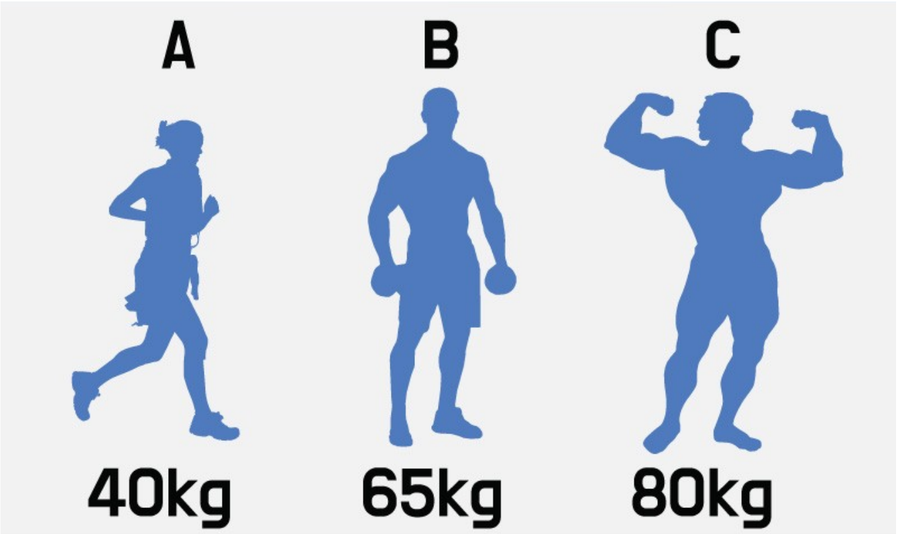
진술 문항	전혀 아니다 <----------		보통이다	----------> 매우 그렇다	
스탯솔은 논문 작성에 필요한 정보를 제공해 준다	1	2	3	4	5
스탯솔은 논문 작성에 필요한 지식을 제공해 준다	1	2	3	4	5
나는 스탯솔에서 많은 도움을 받고 있다	1	2	3	4	5
나는 내 동료에게 스텟솔을 소개시켜 줄 것이다	1	2	3	4	5
나는 스탯솔이 매우 유용한 블로그라 생각한다	1	2	3	4	5

빅데이터를 위한 통계학

통계 기초 개념 2

비율 척도 (ratio scale)

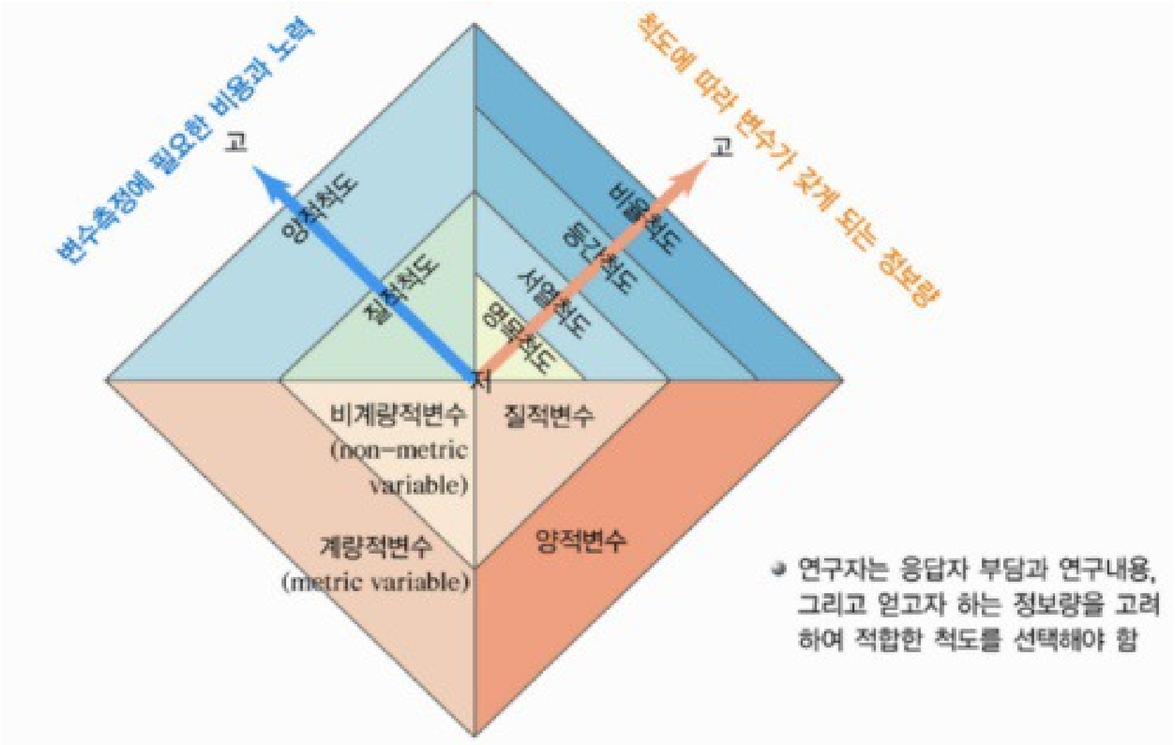
- 등간 척도에 비율의 개념이 추가된 척도로, 절대적 기준값이 존재하는 측정 도구
- 모든 산술적인 사칙연산이 가능함



빅데이터를 위한 통계학

통계 기초 개념 2

척도의 선택 기준



빅데이터를 위한 통계학

통계 기초 개념 2

측정 : 관찰 대상의 특성에 대해 일정한 규칙에 따라 기술적으로 수치를 부여하여 계량화하는 것

- 추론 측정 : 어떤 사물/사건의 속성을 다른 사물/사건의 속성과 관련지어 측정
- 측정
  - 임의 측정 : 어떤 사물의 속성과 측정값 사이에 관계가 있다고 가정하고 측정
  - 본질 측정 : 어떤 사물의 속성을 표현하는 본질적 법칙에 따라 숫자를 부여함

* 더 알고 싶다면?

0011 확률 측도론(Probability and Measure Theory)
통계적 분석에 필요한 실변수 함수의 적분 및 확률론적 해석을 다루기 위한 기본 실수 체계의 특성과 측도 이론을 학습하고, 그 응용으로서의 통계적 해석 방법에 대해 학습한다.

빅데이터를 위한 통계학

통계 기초 개념 2

측도론

◆ def : 확률(probability)과 확률공간(probability space)
다음을 충족시키는 P 를 (Ω,F)위의 확률(probability) 이라고 하고
(Ω, F, P)를 확률공간(Probability Space)이라고 한다.

- $P : F \rightarrow [0,1]$
- $P(\Omega)=1$, $P(A)\in[0,1]$ $\forall A \in F$
- $A_i \cap A_j = \varnothing$ for $i \neq j$ $\Rightarrow P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$

★ 설명충 코너 : 측도와 확률

측도는 집합에 '크기'를 부여하기 위해 만든 개념으로 잴 수 있는 집합들에 실수로 가는 함수를 부여한 것이다.

◆ def. 측도(Measure) 및 측도공간

어떤 집합 X의 부분집합들 중 측도 가능한(Measurable) 집합들의 모임을 M이라 하면 M 에서 $[0,\infty]$로 가는 함수 $\mu : M \rightarrow [0,\infty]$에 대해 다음이 성립

① $\mu(\varnothing)=0$ 공집합의 측도는 0이다!

② $A_i \cap A_j = \varnothing$ for $A_i, A_j \in M$, $i \neq j$ $\Rightarrow \mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$
만나지 않는 집합들의 합집합의 측도는 각 집합의 측도의 합과 같다!
(이를 σ-additive라고 칭함)

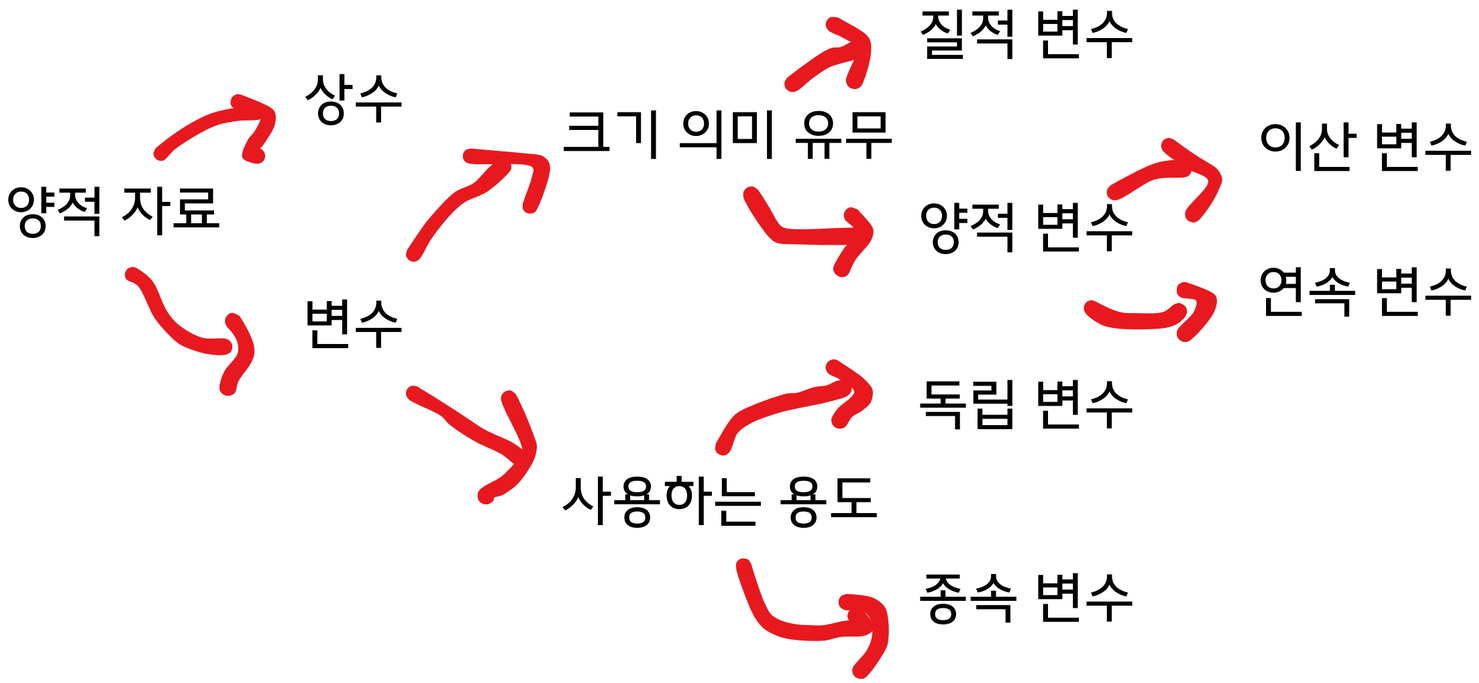
이 경우 (X,M,μ)를 집합 X로 구성된 측도공간(Measure Space)이라고 하며 μ를 M위에서 정의된 측도(Measure)라 한다. 확률(P)은 어떠한 사건(집합 A)의 측도에 해당된다고 할 수 있다.

확률 외에도 많이 쓰이는 측도로는 르벡 측도(Lebesgue Measure)가 있다. 유클리드 공간 $\mathcal{R}^n$에 부여되는 보편적 측도로 길이나 넓이를 수학적으로 엄밀하게 정의한 것이다.

빅데이터를 위한 통계학

통계 기초 개념 2

양적 자료 : 수치로 표현된 자료



빅데이터를 위한 통계학

통계 기초 개념 2

질적 변수 : 비계량적 함수 / non-metric 함수

- 연산이 의미가 없는 변수

ex) 남성 = 1, 여성 = 2

- 주로 명목 척도와 서열 척도를 통해 측정됨

빅데이터를 위한 통계학

통계 기초 개념 2

양적 변수 : 계량적 함수 / metric 함수

- 연산이 가능한 의미 있는 수치로 나타낼 수 있는 변수
ex) 키, 몸무게

- 주로 등간 척도와 비율 척도를 통해 측정됨

빅데이터를 위한 통계학

통계 기초 개념 2

양적 변수 - 이산변수

- 정수값만을 취할 수 있는 변수
- ex) 학생 수, 자동차 판매 대수

빅데이터를 위한 통계학

통계 기초 개념 2

양적 변수 - 연속변수

- 연속적인 모든 실수 값을 취할 수 있는 변수
- ex) 몸무게, 키

빅데이터를 위한 통계학

통계 기초 개념 2

독립 변수 (= 원인, 설명, 예측 변수, feature)

- 독립적으로 변하는 변수
- 조사자에 의해 조작(not cheating)이 가능함

빅데이터를 위한 통계학

통계 기초 개념 2

종속 변수 (= 결과, 목적, 타깃 변수, target)

- 독립 변수에 종속되어 값을 가지는 변수
- 수동적으로 값이 주어짐

빅데이터를 위한 통계학

분포와 특성

분포를 설명하기 위해 필요한 개념

중심성향 평균, 중앙값, 최빈값

산포성향 범위, 평균편차, 분산과 표준편차

왜도와 첨도

빅데이터를 위한 통계학

분포와 특성

중심성향 - 평균 (mean)

- 변수값들을 모두 더한 한계를 그 개수로 나눈 값

(1) 산술평균

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i$$

(2) 가중평균

$$W = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

(3) 기하평균

$$\left(\prod_{i=1}^{n} x_i\right)^n = \sqrt[n]{x_1 x_2 \cdots x_n}$$

(4) 조화평균

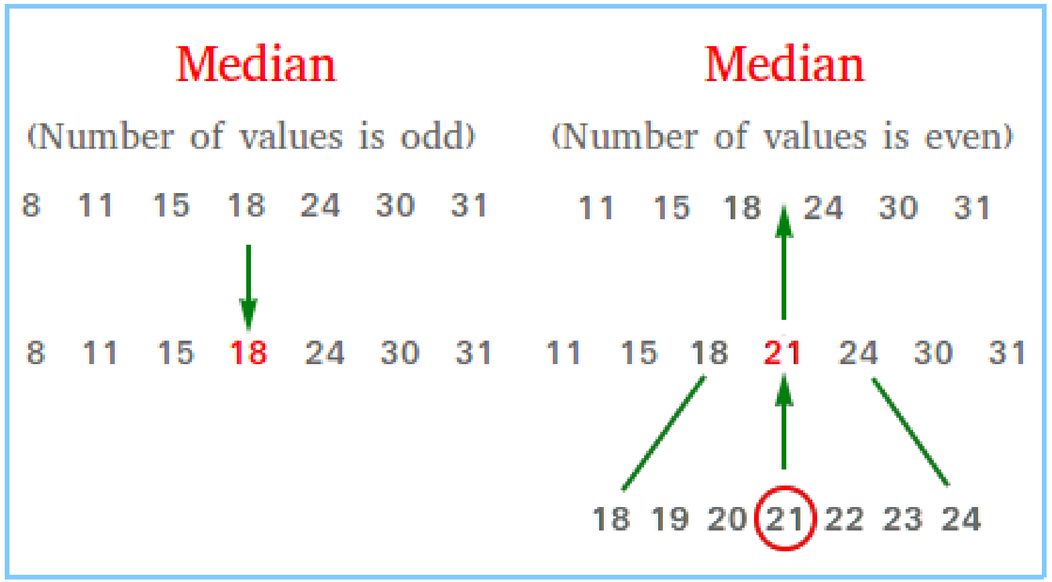
$$H = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\ldots+\frac{1}{x_n}}$$

빅데이터를 위한 통계학

분포와 특성

중심성향 - 중앙값 (median)

- 전체 변수값을 오름이나 내림차순으로 정렬했을 때 중앙에 위치한 값

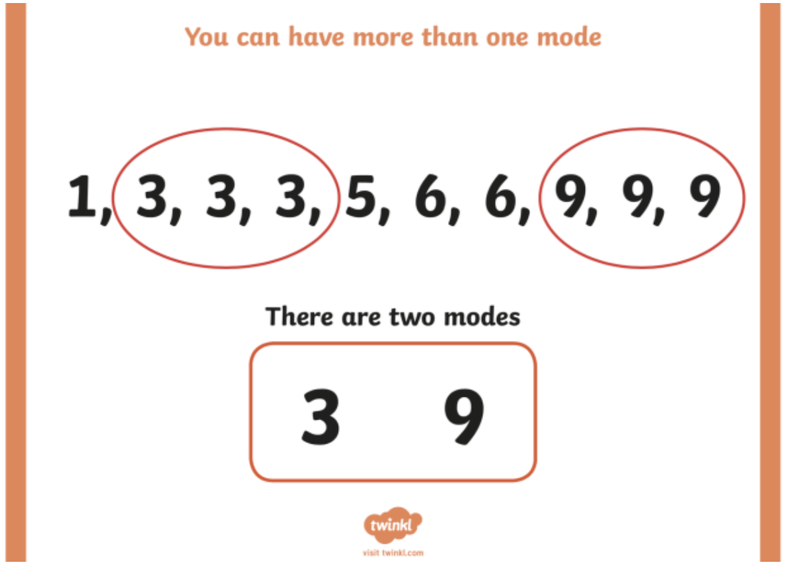


빅데이터를 위한 통계학

분포와 특성

중심성향 - 최빈값 (mode)

- 전체 변수값들의 빈도를 구했을 때 도수가 가장 높은 값



빅데이터를 위한 통계학

분포와 특성

산포성향 - 범위

- 관측치들 중에서 가장 큰 값과 가장 작은 값의 차이
  - 이상치가 있을 경우에는 실제 산포경향을 왜곡시킬 수 있음

빅데이터를 위한 통계학

분포와 특성

산포성향 - 평균편차

- 평균과 개별 관측치 사이 거리의 평균
  - 이상치 문제를 해결함

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)|$$

빅데이터를 위한 통계학

분포와 특성

산포성향 - 분산과 표준편차

- 표준편차는 평균편차보다 정확도는 떨어지지만 더 계산하기 쉬움

분산

$$\sigma^2 = \frac{\sum(xi - \bar{x})^2}{N}$$

표준편차

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

빅데이터를 위한 통계학

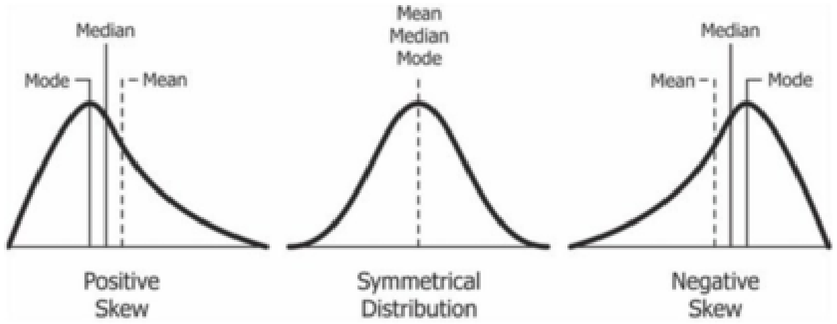
분포와 특성

왜도 : 한쪽으로 기울어진 정도

Pearson's Skewness Coefficient

$$\text{Using Mode: } \frac{\bar{x} - \text{Mode}}{s}$$

$$\text{Using Median: } \frac{3(\bar{x} - \text{Median})}{s}$$

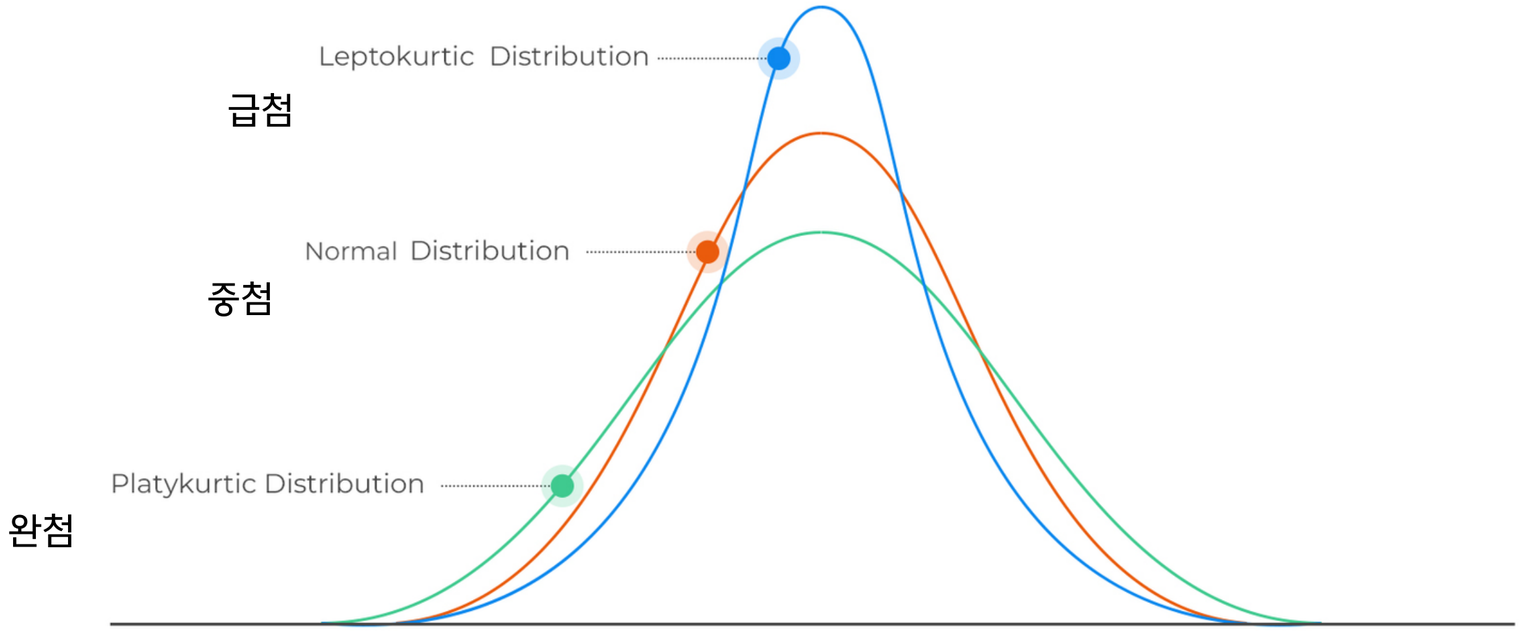


왼쪽꼬리분포 대칭분포 오른쪽꼬리분포

빅데이터를 위한 통계학

분포와 특성

첨도 : 분포가 얼마나 뾰족한지에 대한 정도



빅데이터를 위한 통계학

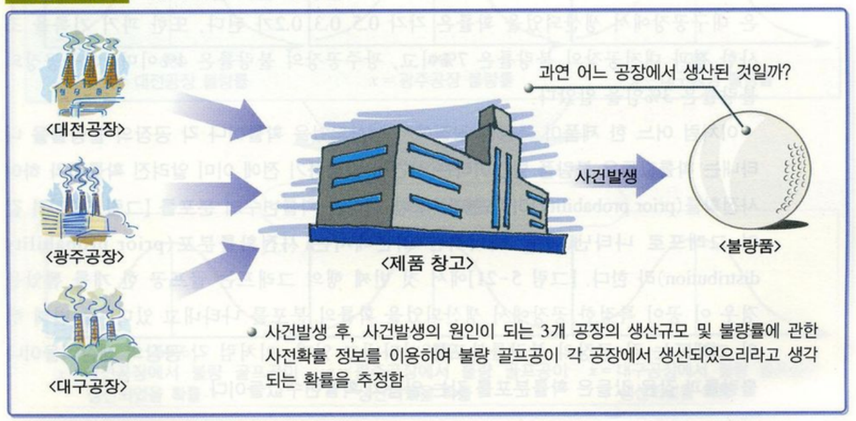
베이지안 이론

베이지안 이론이란

- 주어진 사전확률 정보를 통해 사후확률을 예측함

베이지안 이론이란

그림 5-20 베이지안 이론의 개념



즉 사건이 발생하고 난 후, 사건발생의 원인에 대한 확률(사후)을 사건발생 전에 이미 알고 있는 정보(사전)를 이용하여 구하는 것

빅데이터를 위한 통계학

베이지안 이론

사전 확률과 사후 확률

- 사전 확률 : 이미 알고 있는 정보
- 사후 확률 : 사건 발생의 원인이 되는 확률

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

사전 확률 (prior)

사후 확률 (posterior)

빅데이터를 위한 통계학

베이지안 이론

베이즈 정리

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)}$$

$$= \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B \mid A) \cdot \Pr(A) + \Pr(B \mid A^c) \cdot \Pr(A^c)}.$$

빅데이터를 위한 통계학

베이지안 이론

확장된 베이즈 정리

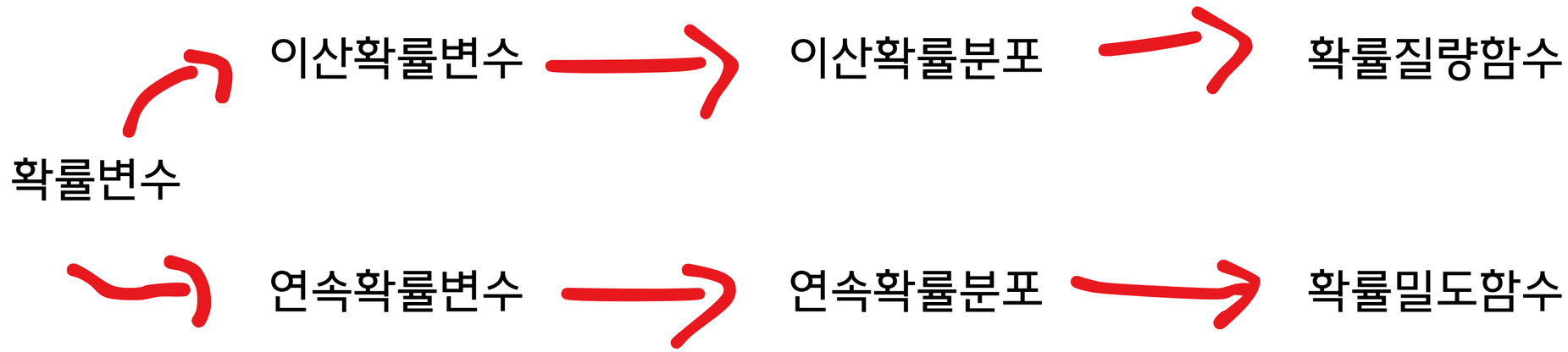
$$P(A|B1,\ldots,Bn) = P(B1,\ldots,Bn|A) * P(A)/P(B1,\ldots,Bn)$$

$$= [P(B1|A)*\ldots*P(Bn|A) * P(A)]/[P(B1|A)*\ldots*P(Bn|A) * P(A) + P(B1|\sim A)*\ldots*P(Bn|\sim A) * P(\sim A)]$$

빅데이터를 위한 통계학

확률분포

확률분포란

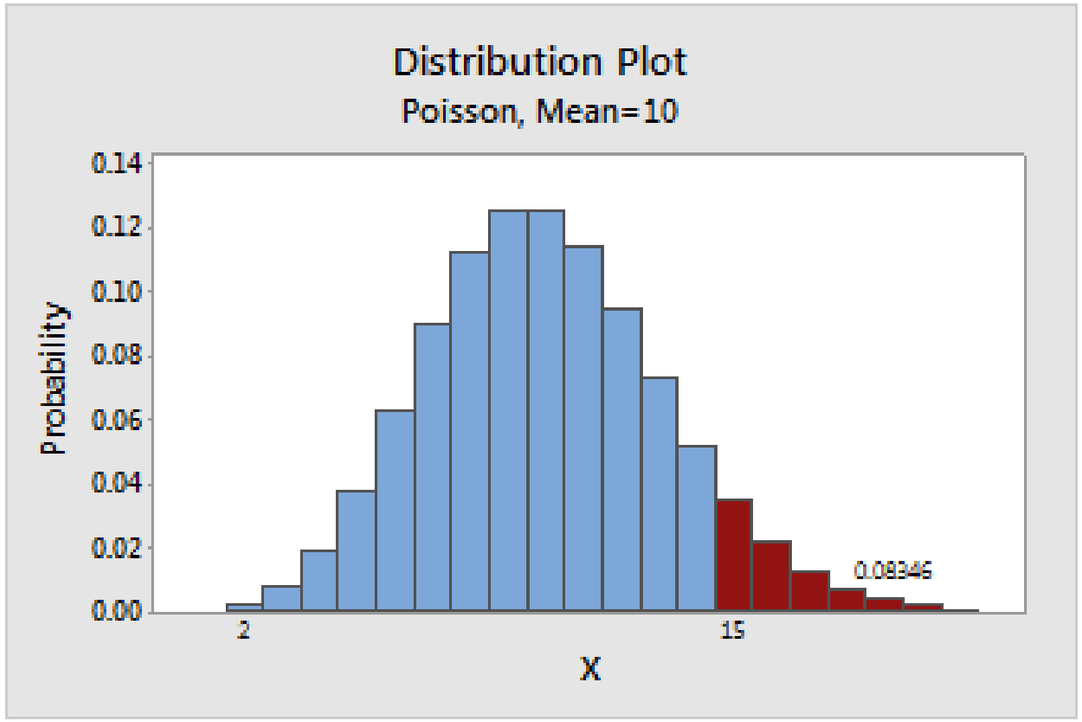


빅데이터를 위한 통계학

확률분포

이산확률분포

- 값이 명확하고 한정적임
- ex)

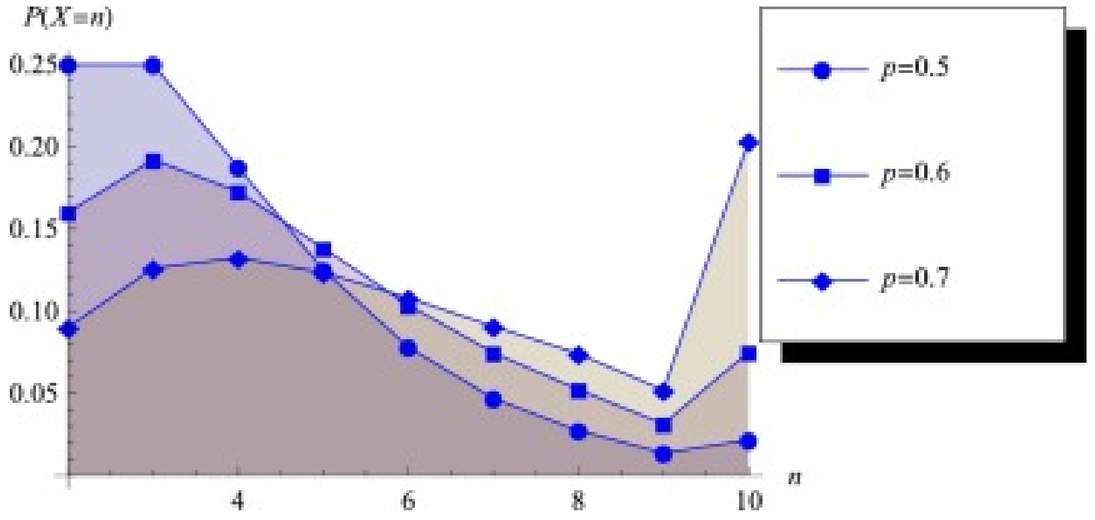


빅데이터를 위한 통계학

확률분포

이산확률분포 - 확률질량함수

- 이산확률분포를 함수로 만든 것

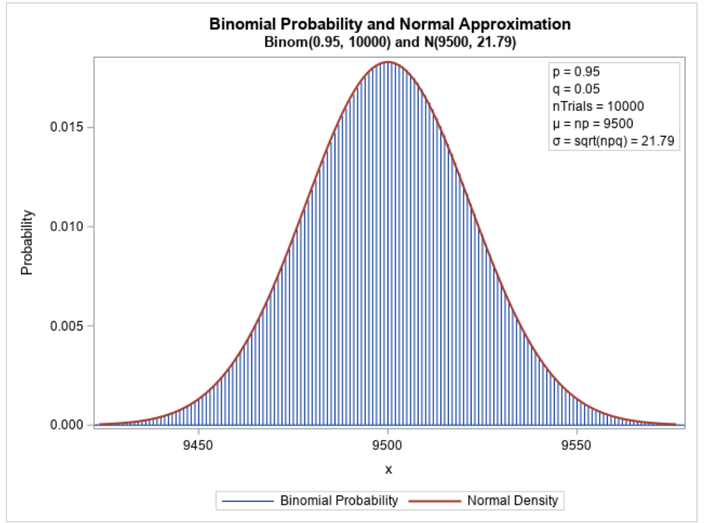


빅데이터를 위한 통계학

확률분포

이산확률분포 - 이항분포

- 베르누이 시행 : 결과값이 2종류인 경우
- 이항분포 : 베르누이 시행 결과의 값을 변수값으로 하는 분포

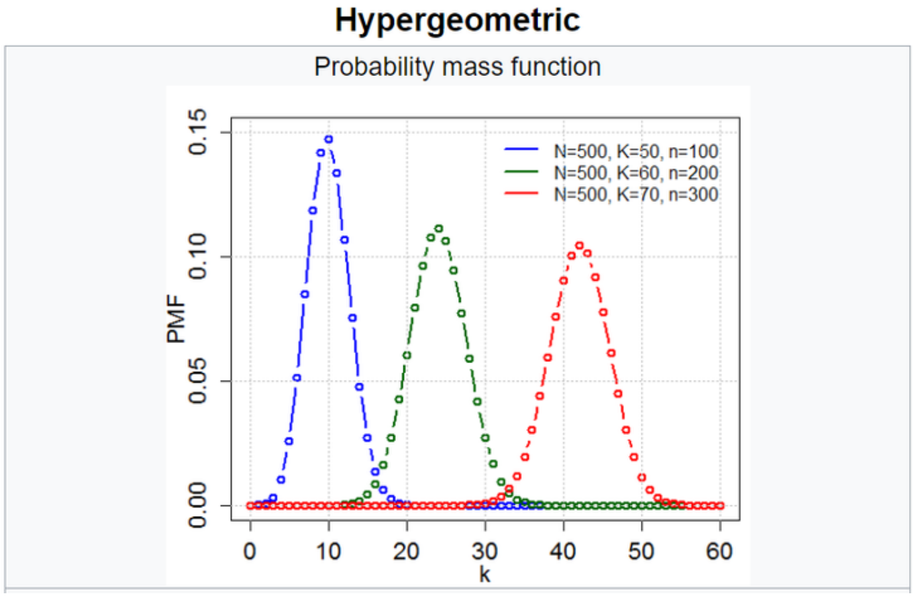


빅데이터를 위한 통계학

확률분포

이산확률분포 - 초기하분포

- 표본공간이 매번 변할 때, 비복원추출을 하는 경우의 확률분포

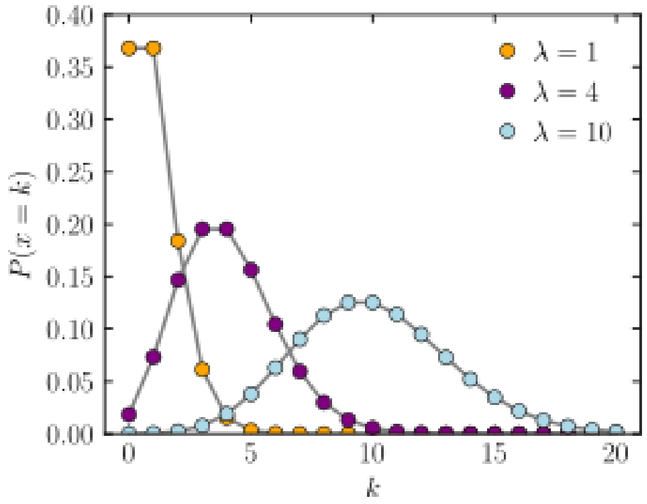


빅데이터를 위한 통계학

확률분포

이산확률분포 - 포아송분포

- 특정 시간 안에 발생하는 특정한 사건의 횟수를 변수값으로 하는 분포

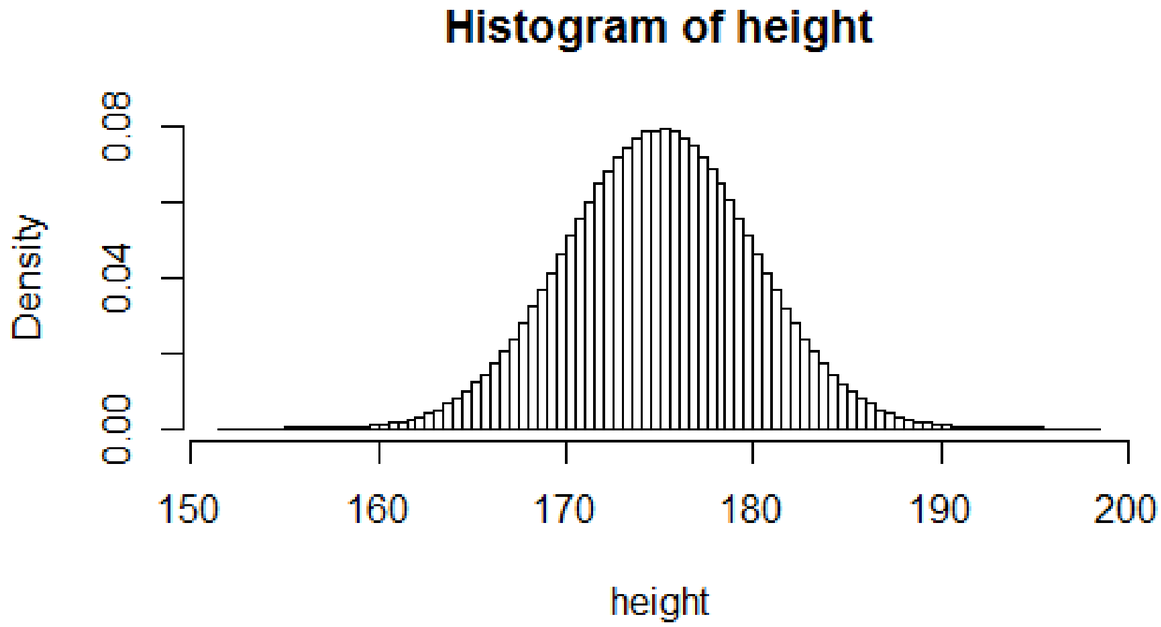


빅데이터를 위한 통계학

확률분포

연속확률분포

- 값이 딱 떨어지지 않고, 개수도 무한대임
- ex)

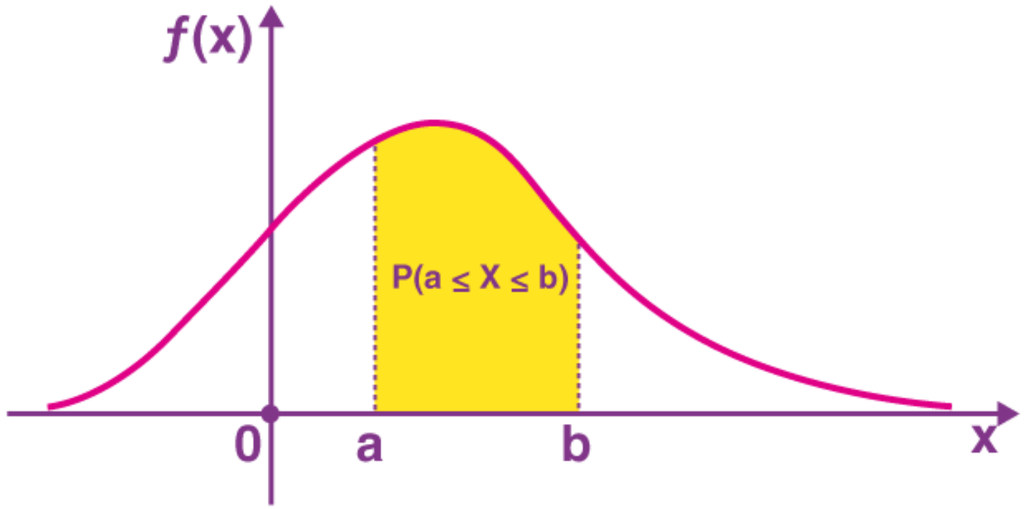


빅데이터를 위한 통계학

확률분포

연속확률분포 - 확률밀도함수

- 연속확률분포를 함수로 만든 것

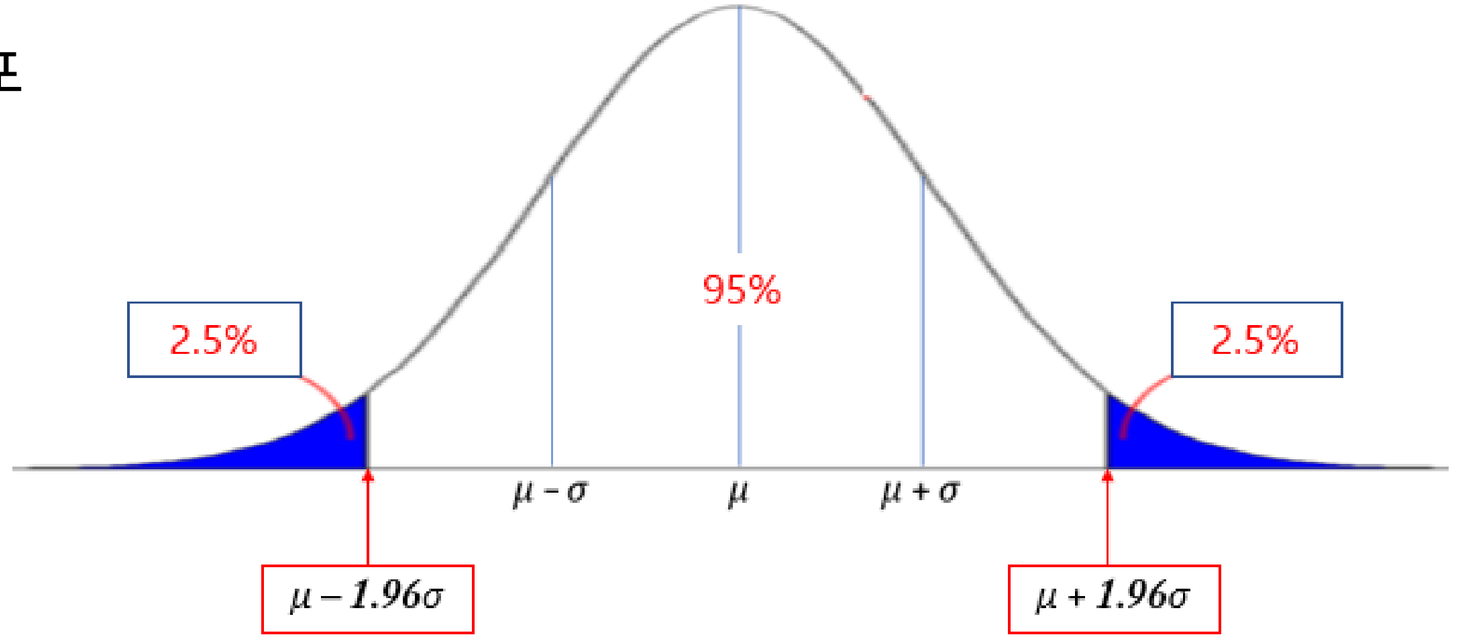


빅데이터를 위한 통계학

확률분포

연속확률분포 - 정규분포

- 종 모양으로 대칭인 분포

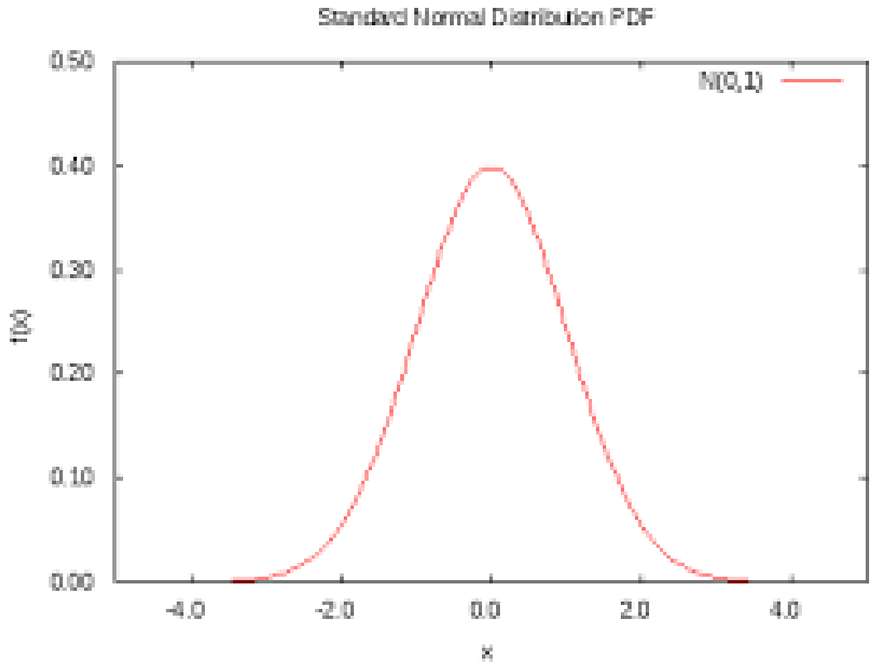


빅데이터를 위한 통계학

확률분포

연속확률분포 - 표준정규분포

- 평균이 0이고 분산이 1인 정규분포

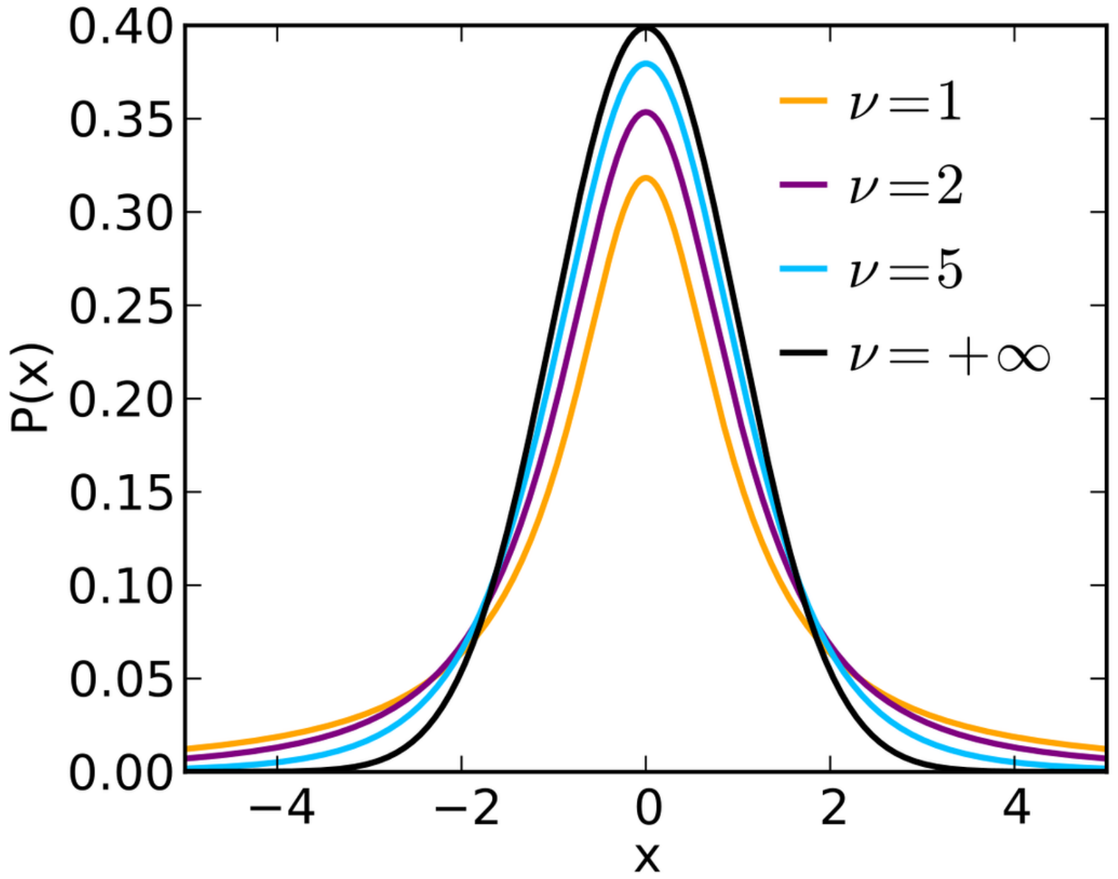


빅데이터를 위한 통계학

확률분포

연속확률분포 - t분포

- 좌우대칭이지만 정규분포보다 완만한 분포

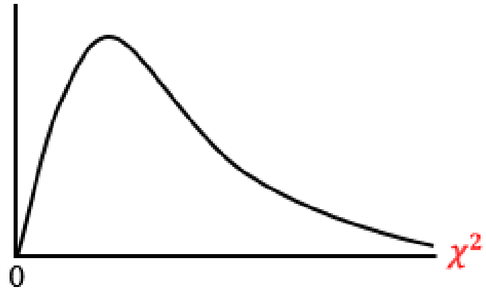


빅데이터를 위한 통계학

확률분포

연속확률분포 - 카이제곱분포

- 분산의 특징을 이용해서 만든 분포



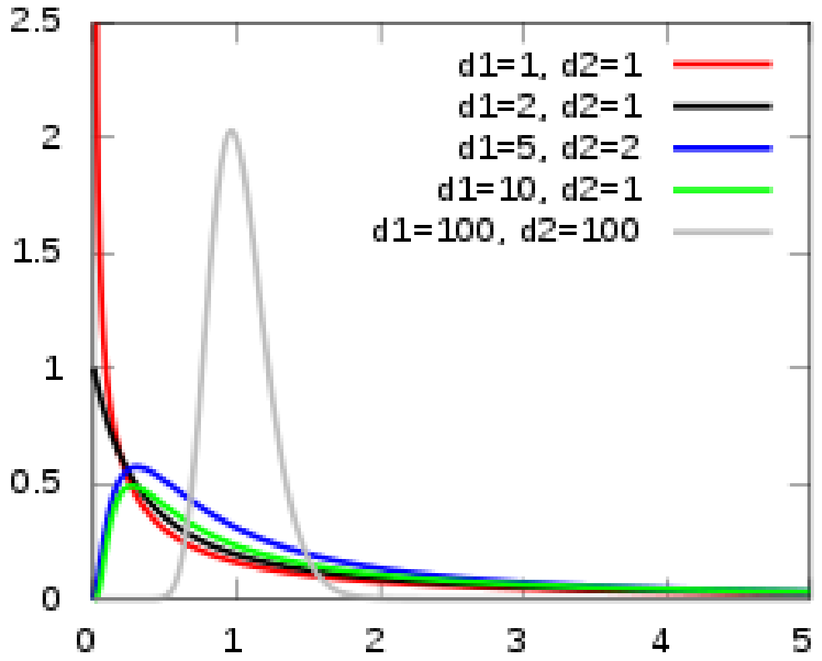
〈 카이제곱분포 〉

빅데이터를 위한 통계학

확률분포

연속확률분포 - F분포

- 카이제곱분포를 이용해서 만든 분포

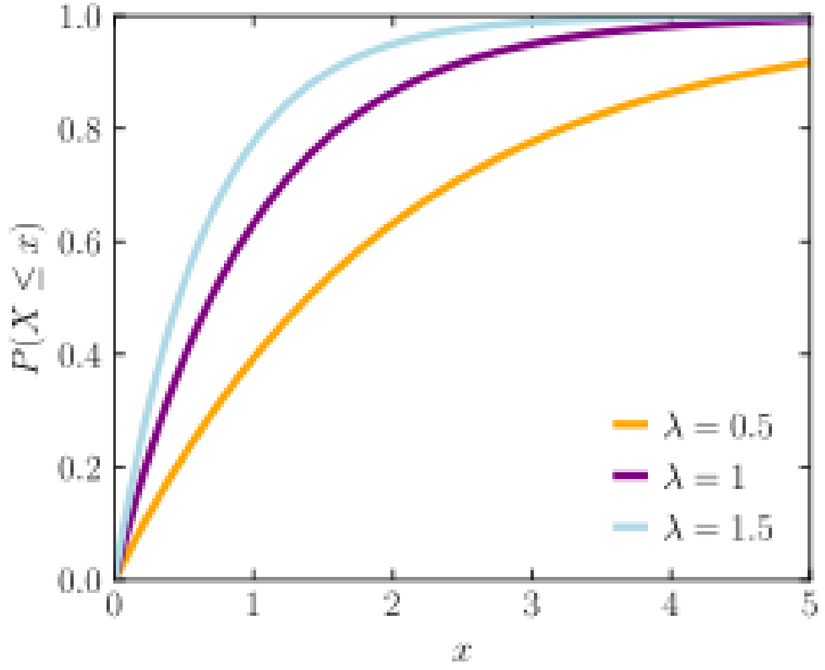


빅데이터를 위한 통계학

확률분포

연속확률분포 - 지수분포

- 사건과 사건 사이의 시간을 변수값으로 하는 분포



수나로움

THANK YOU