

# 통계학과 수학

통계학은 사회, 자연 현상으로 숫자(데이터)와 숫자의 요약(추론)으로 설명하기 위한 학문이다.

통계의 방법론과 추론은 수학을 기반으로 있지만 수학이 물감, 통계학은 아트이다.

# 1. 통계학

## 정의

- [Webster's Dictionary] 통계학(Statistics)은 숫자 데이터(numbered data)를 수집(collect), 정리(summarize), 분석 (analysis), 표현 (presentation)에 관련된 일련의 과정에 관한 학문으로 수학의 한 분야이다.
- [Kendall and Stuart] 통계학은 모집단(population: 관심의 대상이 되는 집단)의 성질(이를 통계학에서는 변수라 한다)을 세거나 측정하여 얻어진 데이터를 다루는 과학의 한 분야이다.
- [Ott] Statistics is about data. 통계학은 데이터에 관한 학문이다.
- [Anonymous] 통계학은 미지에 대한 가이드이다. Statistics is a guide to the unknown.
- [wolfpack] **Statistics is Art.**



### (2) 데이터 정리 (data summarization)

수집된 데이터를 표나 그래프로 (히스토그램, 산점도) 정리하여 그래프 요약 (graphical summary) 얻거나 평균, 표준편차, 비율과 숫자 요약을 얻는다.



### (3) 데이터 분석 (data analysis)

알고자 하는 모집단의 특성 (모수), 함수관계에 대한 추론 단계 추론 (inference)이란 모수 (parameter) 값을 추정하거나 통계적 가설에 대한 진위여부를 판단하는 단계

### (3) 데이터 분석 (data analysis)

알고자 하는 모집단의 특성 (모수), 함수관계에 대한 추론 단계 추론 (inference)이란 모수 (parameter) 값을 추정하거나 통계적 가설에 대한 진위여부를 판단하는 단계

### (4) 데이터 표현 (data presentation)

데이터를 정리, 분석하여 얻은 정보를 적절히 표현하는 단계. 통계 전공자가 아니더라도 쉽게 이해할 수 있도록 작성한다.



## 종류

ASA (American Statistical Association)는 통계학은 데이터 수집, 분석, 표현에 대한 수리적 원리의 과학적 응용이라고 정의하고, 통계학자는 조사 및 실험설계, 데이터 수집, 처리, 분석, 그리고 결과에 대한 해석에 그들의 수리적, 통계학적 지식의 활용하여 과학적 의문에 도움을 주는 전문가이다.

통계학은 기술통계학과 추론통계학으로 나뉜다. 기술통계학은 (descriptive Statistics) 모집단 전체 혹은 표본 (모집단 일부)으로부터 얻은 데이터에 대한 숫자 요약(기초통계량; 평균, 분산)이나 그래프 요약을 (히스토그램) 통하여 관심대상에 대한 유용한 정보를 생산하는 분야이다. 표본으로부터 얻은 정보, 요약된 통계량이나 그래프를 이용하여 모집단의 특성(모수: parameter)을 추론(추정, 검정)하거나 설정된 특성들의 (변수) 함수관계의 진위여부를 판단하는 분야를 추론 통계학이라 (inferential Statistics) 한다.

(a) 기술 통계학(descriptive statistics)

모집단 전체(센서스) 혹은 표본(sample: 모집단 일부)으로부터 얻은 데이터에 대한 숫자 요약(numerical summary, 평균, 분산)이나 그래프 요약(graphical summary)을 통하여 데이터가 가진 정보를 정리한다.

(b) 추론 통계학(inferential statistics)

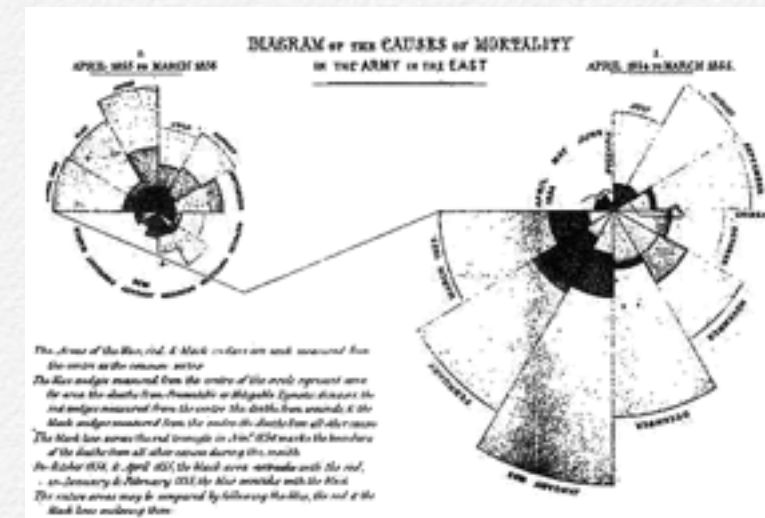
표본으로부터 얻은 정보를 이용하여 모집단의 특성(모수: parameter)을 추론(추정, 검정)하거나 변수들간의 적절한 함수 관계(modeling)를 설정한다.

## 2. 통계 역사

## 1) 기술 통계학 - 국가통계

- 성경: 구약의 민수기(numbers)에 이스라엘 백성 인구 조사, 광야 생활 전과 후에 각 한 번씩 조사

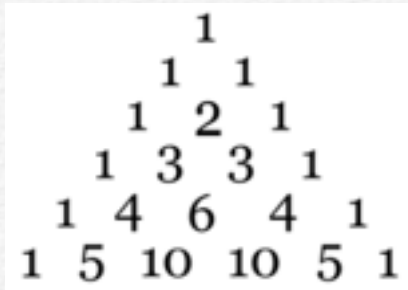
- 로마 황제 Tullis: 세금 징수를 위하여 5년마다 인구 조사, Caesar 가 로마 제국 전역으로 확대
- AD74 로마 패망 후 영국 스페인(미국/페루지역) 센서스
- 센서스(census, 전수 조사): 어원 censura(라틴어 세금tax), statistics(통계학, 라틴어 status국가)
- 17C 영국 출생률과 사망률 조사: 나이팅게일도 통계학자 - polar diagram (원인별 비율)



사회조사는 19세기 후반부터 사회 과학자들에 의해 사용되기 시작하였다. Karl Marx는 1880년에 2만 5천명의 프랑스 노동자를 대상으로 그들의 정치적 태도와 성향에 대한 우편 조사를 시행하였고, Max Weber는 직접 관찰과 조사 연구를 통하여 노동자들의 심리, 태도, 직업관 등을 연구하였다. 20세기에 들어오면서 미국 사회학자들에 의해 사회 조사 방법론 연구가 활발하게 진행되었으며, 미국 통계국(Bureau of Census)은 표본 추출 방법과 자료 수집 방법에 대한 연구에 집중하였고, Gallup과 Roper와 같은 사회여론 조사기관은 방법론을 발전시켰다.

## 2) 추론 통계학 - 확률개념 활용

- 게임 이론(game theory, probability, chance): 중세에 종교적, 도덕적 장벽으로 인하여 발전이 더딤
- Fermat, Pascal(1754): 드멜라라는 친구의 요청으로 2인 게임에서 경기 조합 방법을 계산하기 위하여 파스칼의 삼각형을 제안하였다.



(활용1)

동전을 던져 3번 먼저 이기면 게임이 끝나고 승자는 만원을 가져간다. 동전을 3번 던진 결과 A가 2번 B가 한 번을 이기고 동전을 잃어버려 더 이상 게임을 진행 할 수 없다. 만원을 어떻게 나누어 가져야 하는가?

(활용2)  $(3x + 5y)^5 =$  전개하시오.

- 정규분포 Normal Distribution : 이항실험(동전 던져 앞뒤 나오는 결과 실험)을 무한히 하는 경우 앞면의 회수에 대한 분포 고민(중심극한 정리, DeMoivre, 1667-1754) - Laplace(1749-1827)는 다른 분포의 중심극한 정리

- Gauss, (1777~1855) 정규분포 분포, 행성간 거리 오차에 대한 히스토그램으로부터 식을 유도 - 측정오차의 대부분은 이 분포를 따르고 있어 normal 정규분포라 함. Gaussian 분포
- W.S. Gosset (1908): 독일 양조장 공장장, t-분포, 소표본 평균의 분포가 정규분포에 따르지 않음으로 인하여 발견
- F. Galton(1885): 회귀분석(유전학자, 완두콩과 부모자녀 키의 관계), Karl Pearson 수리적 접근 => 인과관계
- Fisher: 농업 통계 분야 분산분석 방법론 적용

수학적 이론을 바탕으로 통계 방법론이 연구되고 자연 과학은 물론 사회 과학, 인문 과학의 계량적 연구에 이용되고 있음

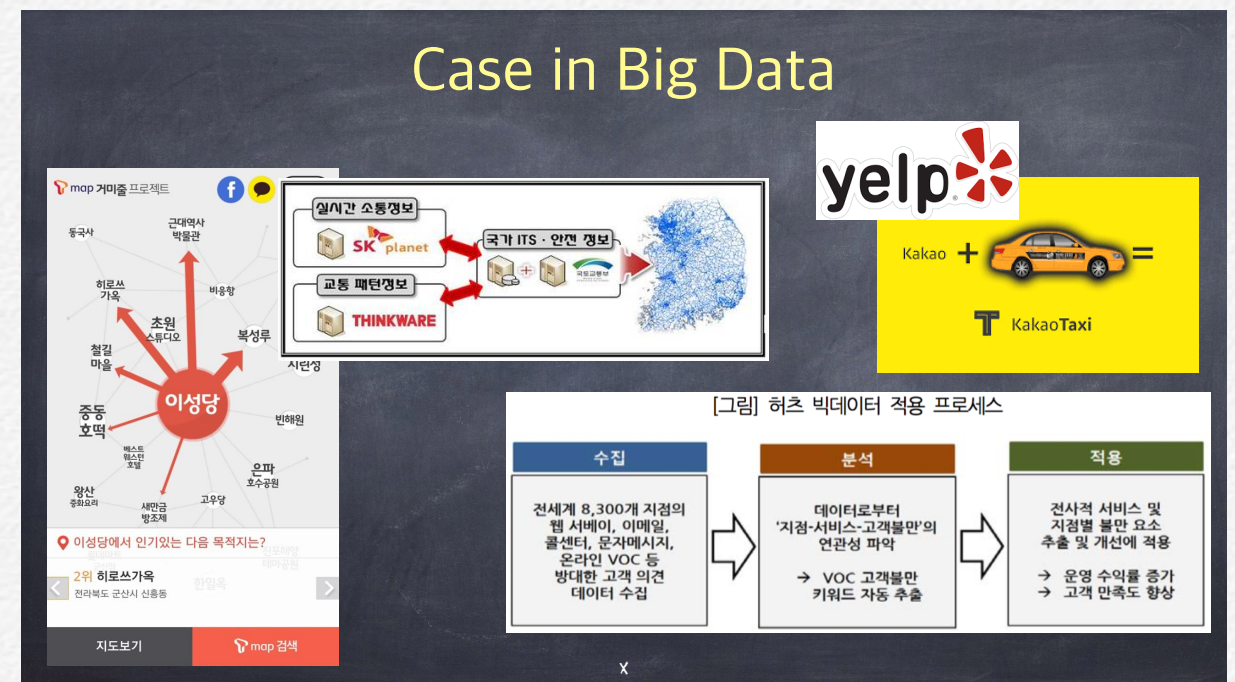
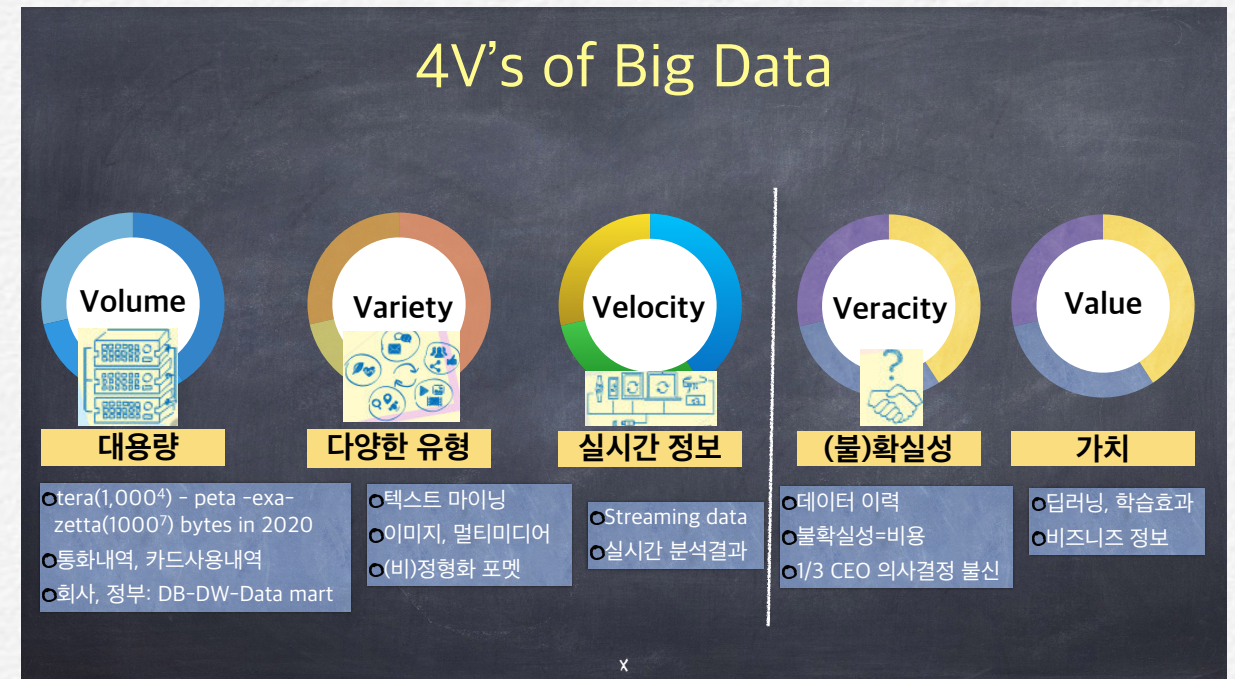
확률 이론 발전 (J. Bernoulli, A. de Moivre, Komogorov), 사회현상에 대한 고찰이나 과학적 사고로 얻어진 논리에 대한 연구와 수학의 접목을 통하여 통계학은 더욱 발전하게 되었다. 천문학자인 Gauss는 행성간 거리를 연구하다가 측정오차의 분포로 정규분포를 (Normal distribution; Gaussian) 유도하였다. 부모의 키와 자녀의 키의 관계를 연구한 유전학자 F. Gallton의 아이디어에 (회귀분석) 통계학자 K. Pearson은 아들 키와 아버지 키의 선형 함수 식의 계수를 계산하는 방법으로 최소자승법을 (method of least square) 제안하였다. 독일의 맥주 공장 공장장인 W. Gosset은 표본의 수가 적을 때는 정규분포를 따르지 않는 경우를 자주 접하게 되어 Student (Gosset의 필명) t-분포를 제안하였다. 사람의 지적 능력을 측정하는 분야를 (Classic, Mathematics, Discovery, English, French, Music) 그룹화 하려는 Gallton의 아이디어에 C. Spearman은 요인분석 방법을 제안하였다. 이처럼 통계학은 인문, 사회, 자연 현상으로부터 발견된 이론을 데이터를 활용하여 설명하거나 진위 여부를 판단하는 방법론과 이론을 제공한다.

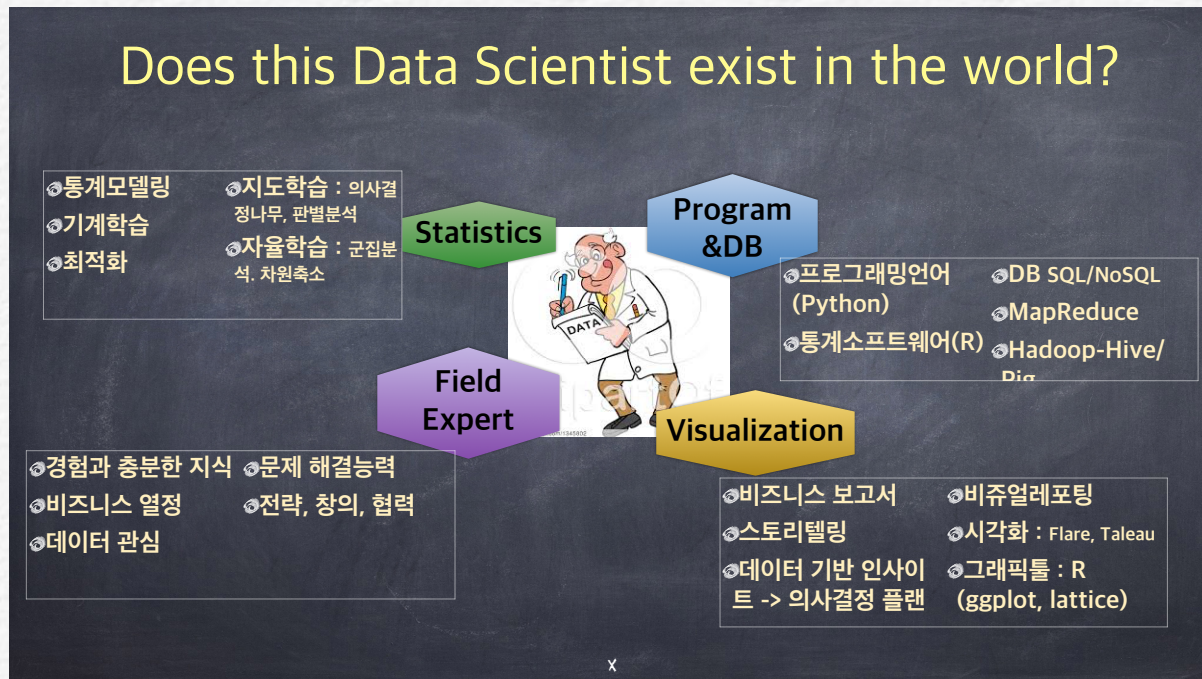
자연과학은 우리 주변 현상을 이해하려는 노력으로 ①현상을 기술하거나 ②일정한 규칙 발견하고 혹은 ③이론과 법칙을 정립하는 과정이다. 현상을 주목하고 경험적인 관찰을 통한 논리적 설명을 한다. 사회과학은 일정한 규칙을 발견하는 것은 쉽지 않으며 인간은 이를 의식적으로 탈피하려는 경향이 있어 일반적인 모형을 적용하는 것에 대한 이견이 적지 않다. 그럼에도 일정한 규칙을 발견하려는 노력은 계속되고 있고 이에 다양한 통계적 방법론이 사용되고 있다. 사회 현상을 이해하고 해석하는데 통계학을 사용할 수 있다는 가능성을 보여준 사람이 벨기에 통계학자 A. Guetelet(19세기 초)이다. 20세기부터 사회 현상에 대한 정확한 서술은 계량화(quantify)라고 생각한 부류와 행동주의(behaviorism) 등이 활발히 연구되었다. 연구과정과 모형화는 경험적 관찰이나 이론을 바탕으로 연구 가설을 정립한다. 다양하고 적절한 방법(관찰, 관측, 실험, 설문조사, 발표자료 수집)으로 데이터를 수집하여 가설의 진위를 판단한다. 가설을 정립할 때 추상적인 개념을 구체화하고 정형화, 수량화(계량화) 작업이 필요하다.

### 3. 통계학과 컴퓨터

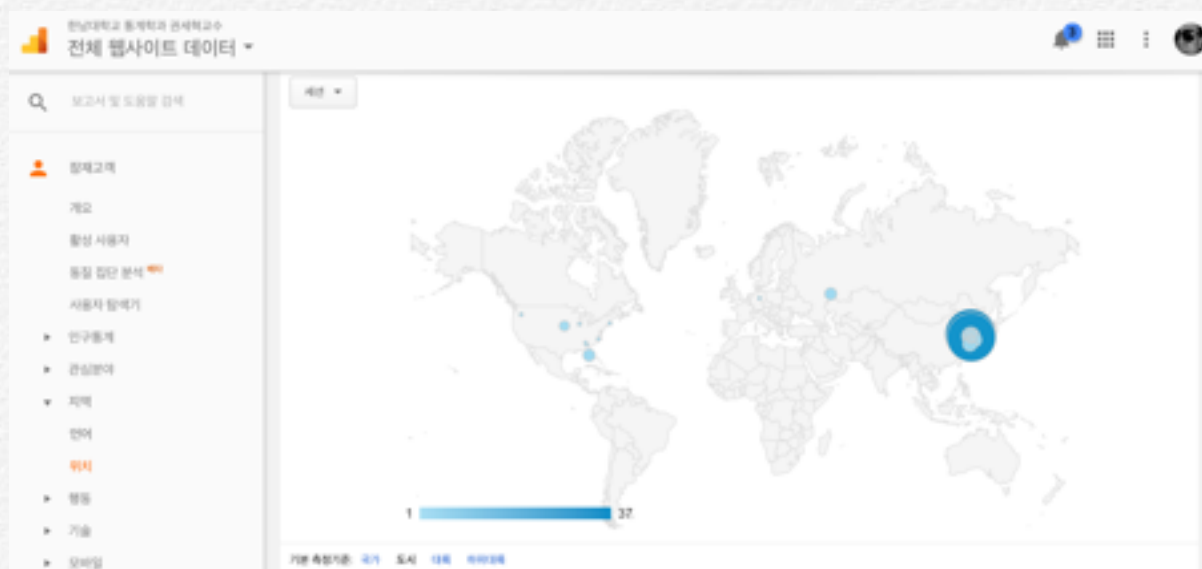
- 통계 소프트웨어 발달로 인하여(SAS, SPSS, Minitab, S-Plus, R) 통계 계산 분석 용이
- 수리적 한계 극복: 복잡한 수학 방정식의 해 혹은 근사 해를 구한다
- Simulation: 사회 현상이나 실험을 컴퓨터에서 실현하여 결과를 예측한다.
- On-Line-Transaction-Protocol ▶ Data warehousing ▶ Data mining ▶ Big Data

### 4. 빅데이터





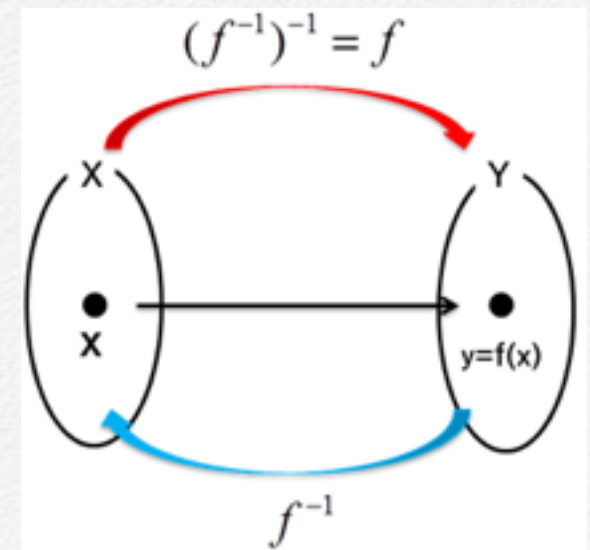
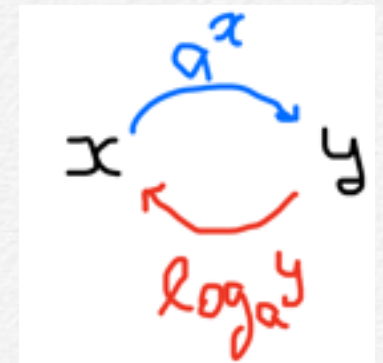
<https://analytics.google.com>



<https://trends.google.com/trends>

## 5. 수학과 통계학

- 통계방법론은 사회, 자연 현상을 설명하기 위해 소프트웨어 발달로 인하여 통계 비전문가라도 데이터로부터 정보(숫자 요약, 그래프, output)를 용이하게 얻을 수 있음
- but, 적절한 분석 방법을 사용하고 올바른 결과 해석을 하려면 통계적 지식이 필요하다.
- 통계학은 숫자의 모임인 데이터로부터 정보를 얻는데 관련된 학문으로 통계적 방법론은 수학 이론 및 전개를 기초로 한다.
- 그러므로 수학적 지식은 통계학 이론을 이해하는데 도움이 되며 자료 분석과 결과 해석을 위한 논리적 사고를 갖게 하므로 통계 전공자가 수리적 사고를 가져야 한다.
- 또한 통계학은 천문학, 물리학의 실험 데이터 분석, 수학의 확률 및 분포 이론으로부터 통계적 방법론이 제안되고 발전되었으므로 통계 전공자에게 수학적 지식은 필요하다.
- 수학은 data scientist가 갖추어야 할 창의력, 논리력을 길러준다. 정보=지식=인터넷인 사회이다.



## Greek 알파벳

|     |       |       |       |        |         |      |         |         |
|-----|-------|-------|-------|--------|---------|------|---------|---------|
| 소문자 | α     | β     | γ     | δ      | ε       | ζ    | η       | θ       |
| 대문자 | A     | B     | Γ     | Δ      | E       | Z    | H       | Θ       |
| 발음  | alpha | beta  | gamma | delta  | epsilon | zeta | eta     | theta   |
| 소문자 | ι     | κ     | χ     | λ      | μ       | ν    | ξ       | ο       |
| 대문자 | I     | K     | X     | Λ      | M       | N    | Ξ       | O       |
| 발음  | iota  | kappa | chi   | lambda | mu      | nu   | xi(ksi) | omicron |
| 소문자 | π     | ρ     | σ     | τ      | υ       | φ    | ψ       | ω       |
| 대문자 | Π     | P     | Σ     | T      | Υ       | Φ    | Ψ       | Ω       |
| 발음  | pi    | rho   | sigma | tau    | upsilon | phi  | psi     | omega   |

## 6. 기초수학

### 지수 Exponent

$a, b$  실수(real number),  $p, q$  양수(positive number),  $m, n$  은 양의 정수(integer) 지수 :  $a^b$  - a=base(밑수), b=exponent(지수)

|                             |                                              |
|-----------------------------|----------------------------------------------|
| (1) $a^0 = 1$               | (2) $a^{-p} = 1 / a^p$                       |
| (3) $a^p a^q = a^{(p+q)}$   | (4) $a^p / a^q = a^{(p-q)}$                  |
| (5) $(a^p)^q = a^{pq}$      | (6) $(ab)^p = a^p b^p$                       |
| (7) $\sqrt[n]{a} = a^{1/n}$ | (8) $\sqrt[n]{ab} = \sqrt[n]{a} \sqrt[n]{b}$ |

(연습) ①  $2^{16} =$                       ②  $\sqrt{9*16} =$                       ③  $9^{2.5} =$

### 로그 Logarithm

- 3의 제곱은 9이다. 세제곱은 27이다. 그럼 얼마 제곱을 해야 10일까?
- 로그  $\log_a b$  - a=base(밑수), b=power(승)
- 지수와 로그는 역함수 개념
- $3^p = 10$ , p는 얼마? - 양변에 로그를 취하면  $\log_{10} 3^p = \log_{10} 10 \Rightarrow p = 1/\log_{10} 3$
- 밑수가 10인 경우 상용 common 로그  $\log_{10}(b) = \log(b)$
- 밑수가 e인 경우 자연 natural 로그  $\log_e(b) = \ln(b)$

|                                       |                                          |
|---------------------------------------|------------------------------------------|
| (1) $\log_a MN = \log_a M + \log_a N$ | (2) $\log_a M / N = \log_a M - \log_a N$ |
| (3) $\log_a M^p = p \log_a M$         | (4) $a^{\log_a b} = b$                   |

(5)  $e^{\ln(x)} = x$

#### 수학상수 e ( Euler’s number)

$e = \lim_{n \rightarrow \infty} (1 + 1/n)^n = 2.71828$  - 통계학 유명분포함수 상수 포함

(연습) ①  $\log_4^{10} =$                       ②  $\log_{10}^{0.4} =$                       ③  $\ln(10) =$

#### 로그와 지수는 역함수

역함수는  $y=x$  기준 대칭

$y = \ln(x),$

$$e^y = x \Rightarrow e^x = y$$

두 함수의 그래프를 그리고  $y=x$  기준 대칭임을 보이시오. (in R)

## Permutation & Combination

### 1) Factorial 계승

a) 자연수  $n : n! = n * (n - 1) * \dots * 1$

b) 양의 실수  $z : z! = \Gamma(z + 1) = \int_0^{\infty} t^z e^{-x} dx$

c) 0.5 자연수 :  $n! = \sqrt{\pi} \prod_{k=0.5}^n k$  (예 :  $2.5 = \sqrt{\pi} * (1/2) * (3/2) * (5/2)$ )

(연습) 1~9까지의 숫자로 3자리 숫자를 만들 수 있는 개수?

2) 순열  $P(n, r) = nPr = \frac{n!}{(n - r)!}$

서로 다른  $n$ 개 원소를  $r$ 개 추출하여 순서대로 열거하는 방법 개수

(연습) 1~9까지의 숫자로 서로 다른 3자리 숫자를 만들 수 있는 개수?

3) 조합  $C(n, r) = nCr = \frac{n!}{r!(n - r)!}$  -> 선택한  $r$ 개 순서나열 개수 회수  $r!$  나누기

서로 다른  $n$ 개 원소를  $r$ 개 추출하여 순서대로 열거하는 방법 개수

(연습) 1~9까지의 숫자로 3개 숫자를 추출할 수 있는 개수?

## 수학 시리즈 Mathematical Series,

### 1) 산술 시리즈 arithmetic series 등차 수열 (d=등차)

$$a + (a + d) + (a + 2d) + \dots + (a + (n - 1)d) \\ = \frac{n}{2}(2a + (n - 1)d)$$

$$3 + 7 + 11 + \dots + 99 =$$

$$2 + 6 + 10 + \dots + 102 =$$

### 2) 기하 시리즈 geometric series 등비수열 (r=등비)

$$a + ar + \dots + ar^{n-1} \quad a + ar + ar^2 + \dots \\ = \frac{a(1 - r^n)}{1 - r} \quad , \quad = \frac{a}{1 - r}, -1 < r < 1$$

$$3 + 9 + 27 + 177147 =$$

$$0.5 + 0.25 + 0.125 + \dots =$$

$$1/2 + 1/6 + 1/18 + \dots =$$

### 3) 이항 시리즈 binomial, Pascal Triangle

$$(a + b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \dots + \binom{n}{n}b^n$$

만약  $a+b=10$ 이면 합이 10이다.

$(3x - 2y)^7$ 에서  $x^4y^3$ 의 계수는 얼마인가?

## 테일러 시리즈

미분가능 함수  $f(x)$ 에 대하여

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)(x - a)^2}{2!} + \frac{f'''(a)(x - a)^3}{3!} + \dots$$

만약  $a = 0$ 인 경우 - Maclaurin series

$$f(x) = f(0) + f'(0)x + \frac{f''(0)x^2}{2!} + \frac{f'''(0)x^3}{3!} + \dots$$

(활용)  $\sqrt{9.1}$  = 근사값 구하기

Show that  $(1 + x)^{-1} = 1 - x + x^2 - x^3 + \dots$

$(1 + x)^{-2} = 1 - 2x + 3x^2 - 4x^3 + \dots$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

## Special Series

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n$$

## 합과 곱의 수학기호

크기가 n개인 데이터 시리즈  $(x_1, x_2, \dots, x_n)$

$$\text{합기호 : } \sum_{i=1}^n x_i \qquad \text{곱기호 : } \prod_{i=1}^n x_i = x_1 \cdot x_2 \cdots x_n$$

$$\sum_{i=0}^5 (1 + i/2) = \qquad \sum_{k=1}^{\infty} 0.75^k =$$

$$\prod_{k=1}^5 (0.75)^k = \qquad \prod_{i=0}^5 (i + 0.5)$$

$x_i = i, \ (i = 0,1,2,\dots,10)$ 인 경우 다음을 구하시오.

$$\sum_{i=1}^3 5i =$$

(1)

$$\sum_{k=1}^3 (3k + 1) =$$

(2)

$$\sum_{i=1}^2 \sum_{k=1}^3 (k + i) =$$

(3)

$$\sum_{i=1}^2 \sum_{k=1}^3 (k + 3) =$$

(4)

$$\sum_{i=1}^2 \sum_{k=1}^3 (i + 3) =$$

(5)

$$\sum_{j=0}^3 (x_j + x_{j+1}) =$$

(6)

$$\sum_{j=0}^2 (x_j x_{j+1}) =$$

(7)

$$\sum_{j=0}^3 2x_{j+2} =$$

(8)

$x_i = i, \ (i = 0,1,2,\dots,10)$ 인 경우 다음을 구하시오.

$$\prod_{i=1}^3 (i + 1)^2 =$$

(1)

$$\prod_{i=1}^3 (i + 1)^i =$$

(2)

$$\prod_{i=0}^{10} x_i =$$

(3)

$$\prod_{i=3}^5 x_i^2 =$$

(4)

$$\prod_{i=1}^3 ix_i =$$

(5)

$$\prod_{i=9}^{10} x_{i+2}x_i =$$

(6)

$$\text{평균 : } E(X) = \frac{\sum_{i=1}^n x_i}{n}, \qquad \text{분산 : } V(X) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n \text{ or } (n - 1)}$$