

# 재미있을 확률이 높은 강연

정준혁 Joonhyuk Jung

May 18, 2024



# About Me



1. 서울과학고등학교 (2014-2017)
  - 26기
2. 서울대학교 (2017-2023)
  - B.S. in Statistics and B.A. in Economics
3. The University of Chicago (2024-)
  - Ph.D. in Statistics

Please visit <https://joonhyuk.com> for more details.

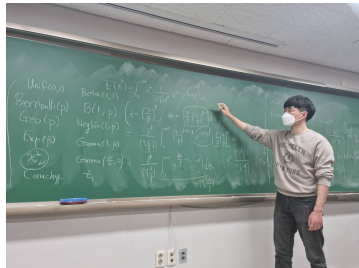


# About Me



1. 서울과학고등학교 (2014-2017)
  - 26기
2. 서울대학교 (2017-2023)
  - B.S. in Statistics and B.A. in Economics
3. The University of Chicago (2024-)
  - Ph.D. in Statistics

Please visit <https://joonhyuk.com> for more details.





1. 통계학이 무엇인지 이해하고, 왜 배워야 하는지 알아본다.



1. 통계학이 무엇인지 이해하고, 왜 배워야 하는지 알아본다.
  - 통계학과 같까요? 수학과 같까요?



1. 통계학이 무엇인지 이해하고, 왜 배워야 하는지 알아본다.
  - 통계학과 같까요? 수학과 같까요?
2. 일반인과 설객인 모두가 만족할 수 있는 설명을 제공한다.



1. 통계학이 무엇인지 이해하고, 왜 배워야 하는지 알아본다.
  - 통계학과 같까요? 수학과 같까요?
2. 일반인과 설객인 모두가 만족할 수 있는 설명을 제공한다.
  - 쉬운 상식부터 어려운 전문지식까지 차근차근

# 통계학이 무엇인가요 (일반인)



## 1. 사전적 정의 (from ChatGPT)



**You**

Define statistics



**ChatGPT**

Statistics is a branch of mathematics that involves collecting, analyzing, interpreting, presenting, and organizing data. It encompasses methods for designing experiments and surveys, as well as techniques for summarizing and drawing conclusions from data. Statistics is widely used in various fields such as science, business, economics, sociology, psychology, and many others to make informed decisions and predictions based on empirical evidence.





# 통계학이 무엇인가요 (일반인)

## 1. 사전적 정의 (from ChatGPT)

- 데이터를 수집, 가공, 분석, 발표하는 수학의 한 분야



You

Define statistics



ChatGPT

Statistics is a branch of mathematics that involves collecting, analyzing, interpreting, presenting, and organizing data. It encompasses methods for designing experiments and surveys, as well as techniques for summarizing and drawing conclusions from data. Statistics is widely used in various fields such as science, business, economics, sociology, psychology, and many others to make informed decisions and predictions based on empirical evidence.



# 통계학이 무엇인가요 (일반인)

## 1. 사전적 정의 (from ChatGPT)

- 데이터를 수집, 가공, 분석, 발표하는 수학의 한 분야
  - 데이터는 또 무엇인가요?



**You**

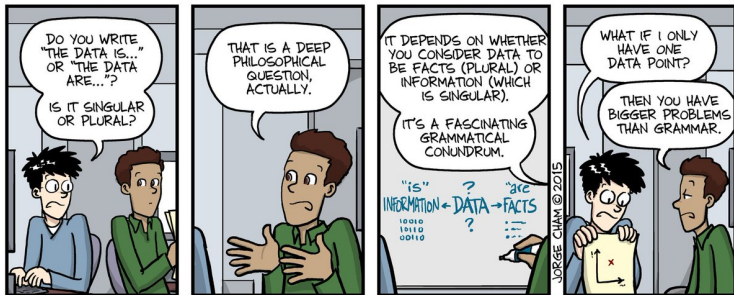
Define statistics



**ChatGPT**

Statistics is a branch of mathematics that involves collecting, analyzing, interpreting, presenting, and organizing data. It encompasses methods for designing experiments and surveys, as well as techniques for summarizing and drawing conclusions from data. Statistics is widely used in various fields such as science, business, economics, sociology, psychology, and many others to make informed decisions and predictions based on empirical evidence.

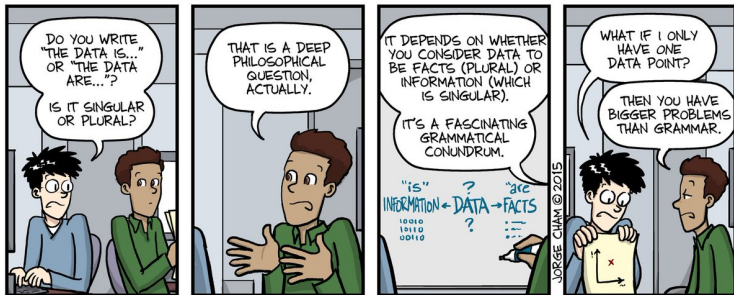
# 데이터가 무엇인가요 (일반인)



WWW.PHDCOMICS.COM

## 1. 영어 data는 영어 명사 datum의 복수형

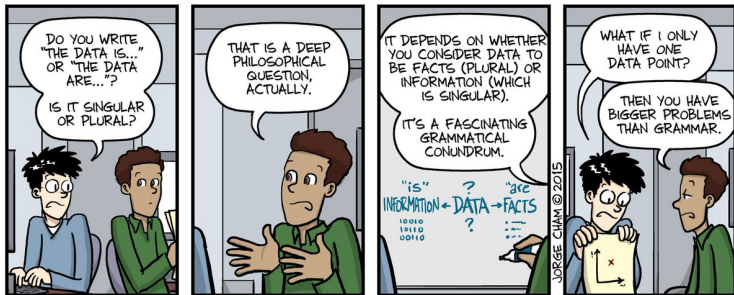
# 데이터가 무엇인가요 (일반인)



WWW.PHDCOMICS.COM

1. 영어 data는 영어 명사 datum의 복수형
2. 라틴어 datum은 라틴어 동사 dare의 과거분사형

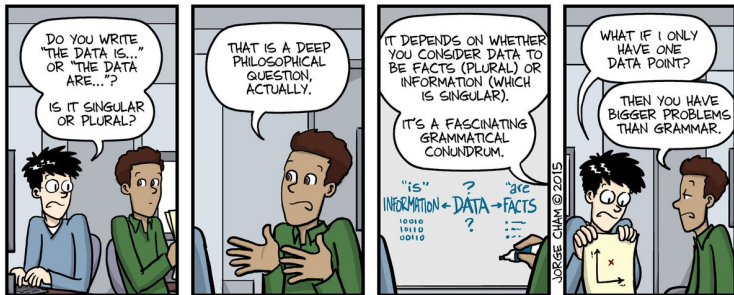
# 데이터가 무엇인가요 (일반인)



WWW.PHDCOMICS.COM

1. 영어 data는 영어 명사 datum의 복수형
2. 라틴어 datum은 라틴어 동사 dare의 과거분사형
  - dare = give = 주다

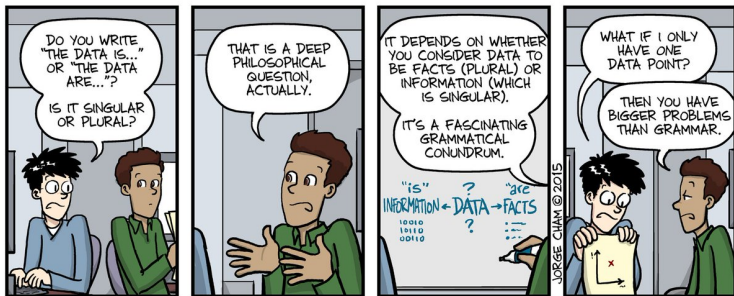
# 데이터가 무엇인가요 (일반인)



WWW.PHDCOMICS.COM

1. 영어 data는 영어 명사 datum의 복수형
2. 라틴어 datum은 라틴어 동사 dare의 과거분사형
  - dare = give = 주다
  - datum = given = 주어진 (것)

# 데이터가 무엇인가요 (일반인)



WWW.PHDCOMICS.COM

1. 영어 data는 영어 명사 datum의 복수형
2. 라틴어 datum은 라틴어 동사 dare의 과거분사형
  - dare = give = 주다
  - datum = given = 주어진 (것)
3. 즉, data는 “주어진 것들”

# 통계학이 무엇인가요 (일반인)

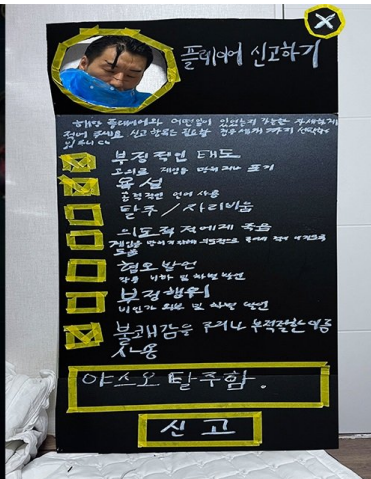


## 1. 더 쉬운 사전적 정의 (from ChatGPT)

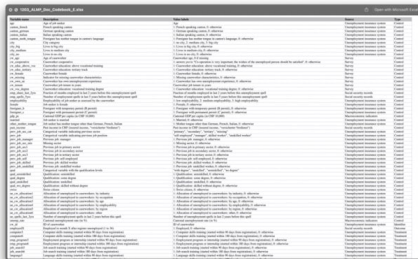
- 주어진 것들을 수집, 가공, 분석, 발표하는 수학의 한 분야
  - 이미 여러분은 직/간접적으로 통계학에 관여하는 중



# 데이터의 수집



# 데이터의 가공, 분석





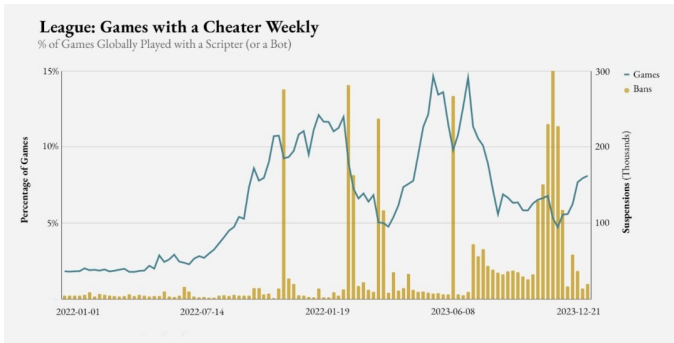
## 언어폭력 행위 신고 피드백

소환사님께서 최근 신고해주신 플레이어를 검토한 결과  
[ 언어폭력 ] 행위가 확인되어 제재하였습니다.

리그 오브 레전드는 소환사님의 신고를 소중히 여기고 검토하겠습니다.  
앞으로도 적극적인 신고로 건전한 게임 문화 양성에 동참해주시기를 부탁드립니다.

확인

# 데이터의 발표



# 통계학을 왜 배워야 하나요



- 단순히 요약하는 것은 진정한 통계학이라고 보기 어렵다.
- 게임사가 사용자 신고(리포트) 분석을 통해 추구하는 것:
  1. 신고 내역을 검토하여 편향, 거짓을 가려냄

# 통계학을 왜 배워야 하나요



- 단순히 요약하는 것은 진정한 통계학이라고 보기 어렵다.
- 게임사가 사용자 신고(리포트) 분석을 통해 추구하는 것:
  1. 신고 내역을 검토하여 편향, 거짓을 가려냄
    - “정말로” 부정적인 영향력을 행사하는 사용자를 제재

# 통계학을 왜 배워야 하나요



- 단순히 요약하는 것은 진정한 통계학이라고 보기 어렵다.
- 게임사가 사용자 신고(리포트) 분석을 통해 추구하는 것:
  1. 신고 내역을 검토하여 편향, 거짓을 가려냄
    - “정말로” 부정적인 영향력을 행사하는 사용자를 제재
  2. 사용자를 분류하거나 미래의 비매너를 예측

# 통계학을 왜 배워야 하나요



- 단순히 요약하는 것은 진정한 통계학이라고 보기 어렵다.
- 게임사가 사용자 신고(리포트) 분석을 통해 추구하는 것:
  1. 신고 내역을 검토하여 편향, 거짓을 가려냄
    - “정말로” 부정적인 영향력을 행사하는 사용자를 제재
  2. 사용자를 분류하거나 미래의 비매너를 예측
    - 신고가 접수되기 전에 불건전한 사용자를 색출



# 통계학을 왜 배워야 하나요



- 단순히 요약하는 것은 진정한 통계학이라고 보기 어렵다.
- 게임사가 사용자 신고(리포트) 분석을 통해 추구하는 것:
  1. 신고 내역을 검토하여 편향, 거짓을 가려냄
    - “정말로” 부정적인 영향력을 행사하는 사용자를 제재
  2. 사용자를 분류하거나 미래의 비매너를 예측
    - 신고가 접수되기 전에 불건전한 사용자를 색출
  3. 사용자의 매너에 관한 가설을 검증

# 통계학을 왜 배워야 하나요



- 단순히 요약하는 것은 진정한 통계학이라고 보기 어렵다.
- 게임사가 사용자 신고(리포트) 분석을 통해 추구하는 것:
  1. 신고 내역을 검토하여 편향, 거짓을 가려냄
    - “정말로” 부정적인 영향력을 행사하는 사용자를 제재
  2. 사용자를 분류하거나 미래의 비매너를 예측
    - 신고가 접수되기 전에 불건전한 사용자를 색출
  3. 사용자의 매너에 관한 가설을 검증
    - 가설: 야스오 사용자는 잔나 사용자보다 비매너가 많다.

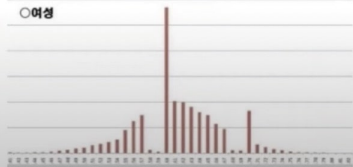
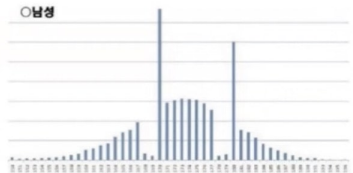
# 통계학을 왜 배워야 하나요 - 편향, 거짓



HIGHEST PAID KOREAN ACTRESS 2021



## 대한민국 성인 키 조사 응답결과



# 통계학을 왜 배워야 하나요 - 분류



# 통계학을 왜 배워야 하나요 - 예측



- 기저귀 옆에 진열할 때 더 잘 팔리는 것은?



# 통계학을 왜 배워야 하나요 - 가설 검정



- 2016년 9월 (고3) 아시아 6등 달성 사냥꾼

# 통계학을 왜 배워야 하나요 - 가설 검정

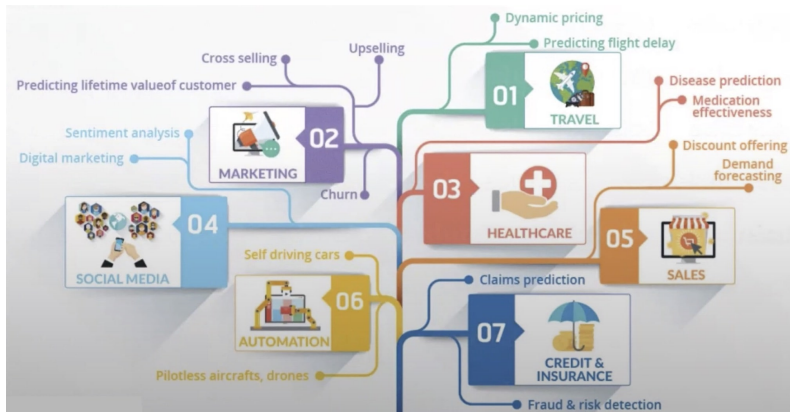


- 크콩이가 나올 확률은 정말  $\frac{1}{3}$  일까?

# 통계학을 왜 배워야 하나요 - 더 나아가서



- 통계학이 적용 가능하지 않은 분야를 찾는 것이 어려움





# 데이터가 무엇인가요 (설곽인)



- 데이터는 단지 **확률변수(random variable)**들의 모임이다.

$$\mathcal{D}_n = (X_1, X_2, \dots, X_n)^\top,$$

$$X_i = \begin{cases} 1, & \text{if } i\text{번째가 쿵쿵이,} \\ 0, & \text{if } i\text{번째가 레오크 혹은 미샤,} \end{cases} \quad i = 1, 2, \dots, n.$$

1. 수집: “야생의 벳”을  $n$ 번 시전한다. 레, 쿵, 미, 쿵, 레, 쿵, ...

# 데이터가 무엇인가요 (설곽인)



- 데이터는 단지 **확률변수(random variable)**들의 모임이다.

$$\mathcal{D}_n = (X_1, X_2, \dots, X_n)^\top,$$

$$X_i = \begin{cases} 1, & \text{if } i\text{번째가 쿵쿵이,} \\ 0, & \text{if } i\text{번째가 레오크 혹은 미샤,} \end{cases} \quad i = 1, 2, \dots, n.$$

1. 수집: “야생의 벳”을  $n$ 번 시전한다. 레, 쿵, 미, 쿵, 레, 쿵, ...
2. 가공: 엑셀 파일 등에 결과를 정리한다. 0, 1, 0, 1, 0, 1, ...

# 데이터가 무엇인가요 (설곽인)



- 데이터는 단지 **확률변수(random variable)**들의 모임이다.

$$\mathcal{D}_n = (X_1, X_2, \dots, X_n)^\top,$$

$$X_i = \begin{cases} 1, & \text{if } i\text{번째가 쿵쿵이,} \\ 0, & \text{if } i\text{번째가 레오크 혹은 미샤,} \end{cases} \quad i = 1, 2, \dots, n.$$

1. 수집: “야생의 벳”을  $n$ 번 시전한다. 레, 쿵, 미, 쿵, 레, 쿵, ...
2. 가공: 엑셀 파일 등에 결과를 정리한다. 0, 1, 0, 1, 0, 1, ...
3. 분석: 표본평균을 계산한다. 일표본 비율검정을 실시한다.

# 데이터가 무엇인가요 (설곽인)



- 데이터는 단지 **확률변수(random variable)**들의 모임이다.

$$\mathcal{D}_n = (X_1, X_2, \dots, X_n)^\top,$$

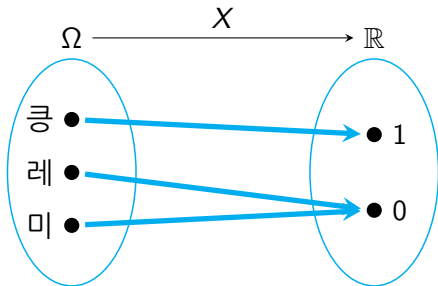
$$X_i = \begin{cases} 1, & \text{if } i\text{번째가 쿵쿵이,} \\ 0, & \text{if } i\text{번째가 레오크 혹은 미샤,} \end{cases} \quad i = 1, 2, \dots, n.$$

1. 수집: “야생의 벳”을  $n$ 번 시전한다. 레, 쿵, 미, 쿵, 레, 쿵, ...
2. 가공: 엑셀 파일 등에 결과를 정리한다. 0, 1, 0, 1, 0, 1, ...
3. 분석: 표본평균을 계산한다. 일표본 비율검정을 실시한다.
4. 발표: “통계학적으로, 쿵쿵이가 나올 확률은  $\frac{1}{3}$ 이 아닙니다.”



# 확률변수가 무엇인가요 (일반인)

1. 무작위적인 상태에 의존하는 값, 변수
2. 각각의 표본에게 실수 값을 부여하는 수학적 규칙, 함수

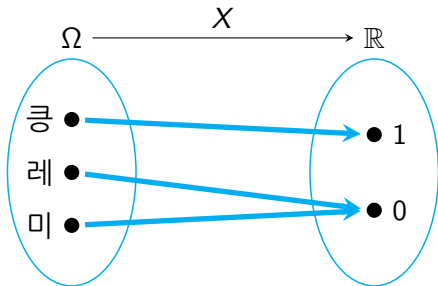


- 가능한 모든 표본  $\omega$ 를 모은 집합  $\Omega$ 를 표본공간이라고 부른다.



# 확률변수가 무엇인가요 (일반인)

1. 무작위적인 상태에 의존하는 값, 변수
2. 각각의 표본에게 실수 값을 부여하는 **수학적인 규칙, 함수**

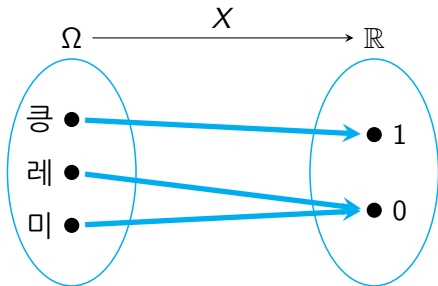


- 가능한 모든 표본  $\omega$ 를 모은 집합  $\Omega$ 를 **표본공간**이라고 부른다.
  - 제비뽑기상자로 비유한다.



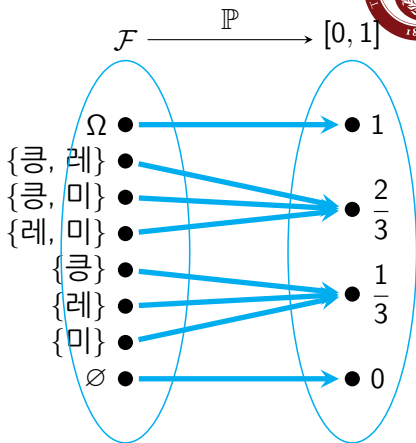
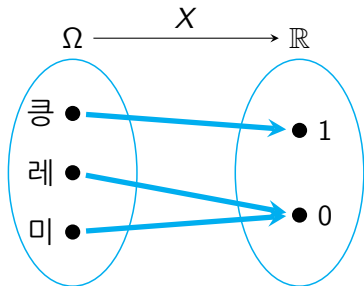
# 확률변수가 무엇인가요 (일반인)

1. 무작위적인 상태에 의존하는 값, 변수
2. 각각의 표본에게 실수 값을 부여하는 **수학적인 규칙, 함수**



- 가능한 모든 표본  $\omega$ 를 모은 집합  $\Omega$ 를 **표본공간**이라고 부른다.
  - 제비뽑기상자로 비유한다.
- 그런데  $X = 1$ 일 **확률**이라는 말은 무슨 뜻일까?  $\mathbb{P}(X = 1) = ?$

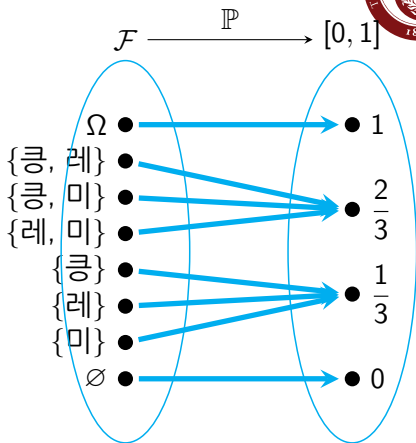
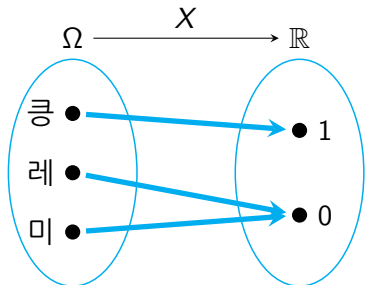
# 확률변수가 무엇인가요 (설곽인)



- 엄밀히 말하면, 확률  $P : \mathcal{F} \rightarrow [0, 1]$ 이 먼저 주어질 때 확률변수  $X$ 를 논할 수가 있다.



# 확률변수가 무엇인가요 (설곽인)



- 엄밀히 말하면, 확률  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ 이 먼저 주어질 때 확률변수  $X$ 를 논할 수가 있다.
- $A \subset \mathbb{R}$ 에 대하여  $\mathbb{P}(X \in A) := \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\})$

# 확률변수가 무엇인가요 (전공자)



제비뽑기상자, 즉 표본공간을 닫힌 구간  $\Omega = [0, 1]$ 로 확장하자.

- Giuseppe Vitali(1875-1932)는 다음과 같은 집합  $V$ 를 발견



# 확률변수가 무엇인가요 (전공자)



제비뽑기상자, 즉 표본공간을 달린 구간  $\Omega = [0, 1]$ 로 확장하자.

- Giuseppe Vitali(1875-1932)는 다음과 같은 집합  $V$ 를 발견
  - $V \subset \Omega = [0, 1]$



# 확률변수가 무엇인가요 (전공자)



제비뽑기상자, 즉 표본공간을 닫힌 구간  $\Omega = [0, 1]$ 로 확장하자.

- Giuseppe Vitali(1875-1932)는 다음과 같은 집합  $V$ 를 발견
  - $V \subset \Omega = [0, 1]$
  - 모든  $r \in \mathbb{R}$ 에 대하여  $r - v \in \mathbb{Q}$ 인  $v \in V$ 가 유일하게 존재



# 확률변수가 무엇인가요 (전공자)



제비뽑기상자, 즉 표본공간을 달린 구간  $\Omega = [0, 1]$ 로 확장하자.

- Giuseppe Vitali(1875-1932)는 다음과 같은 집합  $V$ 를 발견
  - $V \subset \Omega = [0, 1]$
  - 모든  $r \in \mathbb{R}$ 에 대하여  $r - v \in \mathbb{Q}$ 인  $v \in V$ 가 유일하게 존재
- 유일하다는 것은 countable collection  $(q + V)_{q \in \mathbb{Q}}$ 가 disjoint



# 확률변수가 무엇인가요 (전공자)



제비뽑기상자, 즉 표본공간을 닫힌 구간  $\Omega = [0, 1]$ 로 확장하자.

- Giuseppe Vitali(1875-1932)는 다음과 같은 집합  $V$ 를 발견
  - $V \subset \Omega = [0, 1]$
  - 모든  $r \in \mathbb{R}$ 에 대하여  $r - v \in \mathbb{Q}$ 인  $v \in V$ 가 유일하게 존재
- 유일하다는 것은 countable collection  $(q + V)_{q \in \mathbb{Q}}$ 가 disjoint
- 존재한다는 것은 countable union  $\bigcup_{q \in \mathbb{Q}} (q + V) = \mathbb{R}$



# 확률변수가 무엇인가요 (전공자)



- $\Omega$ 에서 아무 표본이나 뽑았을 때 유리수일 확률은 분명 0이다.

# 확률변수가 무엇인가요 (전공자)



- $\Omega$ 에서 아무 표본이나 뽑았을 때 유리수일 확률은 분명 0이다.
- 그럼  $V$ 에 속할 확률은 얼마일까?



# 확률변수가 무엇인가요 (전공자)



- $\Omega$ 에서 아무 표본이나 뽑았을 때 유리수일 확률은 분명 0이다.
- 그럼  $V$ 에 속할 확률은 얼마일까?
- 혹은, 함수  $X_V : \Omega \rightarrow \mathbb{R}$ 을

$$X_V(\omega) = \begin{cases} 1, & \text{if } \omega \in V, \\ 0, & \text{if } \omega \notin V. \end{cases}$$

와 같이 정의할 때  $\mathbb{P}(X_V = 1)$ 은 얼마일까?

# 확률변수가 무엇인가요 (전공자)



- $\Omega$ 에서 아무 표본이나 뽑았을 때 유리수일 확률은 분명 0이다.
- 그럼  $V$ 에 속할 확률은 얼마일까?
- 혹은, 함수  $X_V : \Omega \rightarrow \mathbb{R}$ 을

$$X_V(\omega) = \begin{cases} 1, & \text{if } \omega \in V, \\ 0, & \text{if } \omega \notin V. \end{cases}$$

와 같이 정의할 때  $\mathbb{P}(X_V = 1)$ 은 얼마일까?

- $\mathbb{P}(X_V = 1) = \mathbb{P}(\{\omega \in \Omega : X_V(\omega) = 1\}) = \mathbb{P}(V)$ 가 되는데...

# 확률변수가 무엇인가요 (전공자)



- $\Omega$ 에서 아무 표본이나 뽑았을 때 유리수일 확률은 분명 0이다.
- 그럼  $V$ 에 속할 확률은 얼마일까?
- 혹은, 함수  $X_V : \Omega \rightarrow \mathbb{R}$ 을

$$X_V(\omega) = \begin{cases} 1, & \text{if } \omega \in V, \\ 0, & \text{if } \omega \notin V. \end{cases}$$

와 같이 정의할 때  $\mathbb{P}(X_V = 1)$ 은 얼마일까?

- $\mathbb{P}(X_V = 1) = \mathbb{P}(\{\omega \in \Omega : X_V(\omega) = 1\}) = \mathbb{P}(V)$ 가 되는데...
- $\mathbb{P}(V)$ 의 존재 자체가 확률의 공리계에 모순

# 확률변수가 무엇인가요 (전공자)



- $\Omega$ 에서 아무 표본이나 뽑았을 때 유리수일 확률은 분명 0이다.
- 그럼  $V$ 에 속할 확률은 얼마일까?
- 혹은, 함수  $X_V : \Omega \rightarrow \mathbb{R}$ 을

$$X_V(\omega) = \begin{cases} 1, & \text{if } \omega \in V, \\ 0, & \text{if } \omega \notin V. \end{cases}$$

와 같이 정의할 때  $\mathbb{P}(X_V = 1)$ 은 얼마일까?

- $\mathbb{P}(X_V = 1) = \mathbb{P}(\{\omega \in \Omega : X_V(\omega) = 1\}) = \mathbb{P}(V)$ 가 되는데...
- $\mathbb{P}(V)$ 의 존재 자체가 확률의 공리계에 모순
  - $\mathbb{P}(V)$ 는 정의하지 않기로 함

# 확률변수가 무엇인가요 (전공자)



- $\Omega$ 에서 아무 표본이나 뽑았을 때 유리수일 확률은 분명 0이다.
- 그럼  $V$ 에 속할 확률은 얼마일까?
- 혹은, 함수  $X_V : \Omega \rightarrow \mathbb{R}$ 을

$$X_V(\omega) = \begin{cases} 1, & \text{if } \omega \in V, \\ 0, & \text{if } \omega \notin V. \end{cases}$$

와 같이 정의할 때  $\mathbb{P}(X_V = 1)$ 은 얼마일까?

- $\mathbb{P}(X_V = 1) = \mathbb{P}(\{\omega \in \Omega : X_V(\omega) = 1\}) = \mathbb{P}(V)$ 가 되는데...
- $\mathbb{P}(V)$ 의 존재 자체가 확률의 공리계에 모순
  - $\mathbb{P}(V)$ 는 정의하지 않기로 함
  - $X_V$ 는 확률변수라고 부르지 않기로 함

# 확률변수가 무엇인가요 (전공자)



- $\mathbb{P}(V)$ 의 존재 자체가 불가능

# 확률변수가 무엇인가요 (전공자)



- $\mathbb{P}(V)$ 의 존재 자체가 불가능
- 확률  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ 의 정의역은  $\mathcal{F}$

# 확률변수가 무엇인가요 (전공자)



- $\mathbb{P}(V)$ 의 존재 자체가 불가능
- 확률  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ 의 정의역은  $\mathcal{F}$
- $V \in \mathcal{F}$ 를 고집하면 확률을 정의 못 내릴 수도 있다.



# 확률변수가 무엇인가요 (전공자)



- $\mathbb{P}(V)$ 의 존재 자체가 불가능
- **확률**  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ 의 정의역은  $\mathcal{F}$
- $V \in \mathcal{F}$ 를 고집하면 **확률**을 정의 못 내릴 수도 있다.
- 버릴 것들 버리고, 조금만 골라담자.  $V \notin \mathcal{F} \subsetneq 2^\Omega$

# 확률변수가 무엇인가요 (전공자)



- $\mathbb{P}(V)$ 의 존재 자체가 불가능
- **확률**  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ 의 정의역은  $\mathcal{F}$
- $V \in \mathcal{F}$ 를 고집하면 **확률**을 정의 못 내릴 수도 있다.
- 버릴 것들 버리고, 조금만 골라담자.  $V \notin \mathcal{F} \subsetneq 2^\Omega$
- 남은  $\mathcal{F}$ 의 원소 각각을 **사건**이라고 부른다.

# 확률변수가 무엇인가요 (전공자)



- $\mathbb{P}(V)$ 의 존재 자체가 불가능
- **확률**  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ 의 정의역은  $\mathcal{F}$
- $V \in \mathcal{F}$ 를 고집하면 **확률**을 정의 못 내릴 수도 있다.
- 버릴 것들 버리고, 조금만 골라담자.  $V \notin \mathcal{F} \subsetneq 2^\Omega$
- 남은  $\mathcal{F}$ 의 원소 각각을 **사건**이라고 부른다.
  - Vitali의 집합  $V$  같은 것은 통계학에서 다루는 **사건**이 아니다.

# 확률변수가 무엇인가요 (전공자)



- $\mathbb{P}(V)$ 의 존재 자체가 불가능
- **확률**  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ 의 정의역은  $\mathcal{F}$
- $V \in \mathcal{F}$ 를 고집하면 **확률**을 정의 못 내릴 수도 있다.
- 버릴 것들 버리고, 조금만 골라담자.  $V \notin \mathcal{F} \subsetneq 2^\Omega$
- 남은  $\mathcal{F}$ 의 원소 각각을 **사건**이라고 부른다.
  - Vitali의 집합  $V$  같은 것은 통계학에서 다루는 **사건**이 아니다.
- 기초통계학 시간에 **사건**을 엄밀하게 정의하기 어려운 까닭을 생각해보자.

# 확률변수가 무엇인가요 (전공자)



- 사건공간  $\mathcal{F}$ 가 주어질 때,
- 모든  $x \in \mathbb{R}$ 에 대하여  $X \leq x$ 가 사건이 되도록 하는,
- 보다 엄밀하게는,

$$(X \leq x) := \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}, \forall x \in \mathbb{R}$$

을 성립시키는 함수  $X : \Omega \rightarrow \mathbb{R}$ 을 확률변수라고 부른다.

- 함수  $X_V$  같은 것은 통계학에서 다루는 확률변수가 아니다.

# 통계학과 수학의 비교



- 사건, 확률, 확률변수 등을 정의하려면 대학교 4학년 이상의 수학이 필요하다.
- 하지만 기초통계학 시간에 여러 사건의 확률을 이미 다루었다!
- 통계학자의 언어는 수학자의 언어와 다르다.
- 통계학자의 언어에는 더 많은 직관이 담겨있다.

# 통계학과 수학의 비교 (확률론 전공자)



- 왼쪽은 고2 수준이고 오른쪽은 대학교 4학년 이상 수준!

통계학	수학 (실해석학)
표본공간 $\Omega$	전체집합 $\Omega$
표본 $\omega \in \Omega$	원소 $\omega \in \Omega$
사건공간 $\mathcal{F}$	$\sigma$ -대수 $\mathcal{M}$
사건 $E \in \mathcal{F}$	가측집합 $E \in \mathcal{M}$
확률 $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$	(유한)측도 $\mu : \mathcal{M} \rightarrow [0, 1]$
$\mathbb{P}(E)$	$\mu(E) = \int_E d\mu$
$\mathbb{P}(\Omega) = 1$	$\mu(\Omega) = \int d\mu = 1$
확률변수 $X : \Omega \rightarrow \mathbb{R}$	가측함수 $g : \Omega \rightarrow \mathbb{R}$

# 통계학과 수학의 비교 (확률론 전공자)



통계학	수학
확률변수 $X : \Omega \rightarrow \mathbb{R}$ X의 확률분포	가측함수 $g : \Omega \rightarrow \mathbb{R}$ 보렐측도 $(\mu \circ g^{-1}) : \mathcal{B}_{\mathbb{R}} \rightarrow [0, 1]$
$h(X)$ 의 평균, 기댓값 $\mathbb{E}(h(X))$	합성함수 $(h \circ g)$ 의 르벡적분 $\int (h \circ g) d\mu = \int h d(\mu \circ g^{-1})$
X의 누적분포함수 $F_X(t) = \mathbb{P}(X \leq t)$	보렐측도 $(\mu \circ g^{-1})$ 의 누적함수 $F_g(t) = \mu(g^{-1}(-\infty, t])$
X의 확률밀도함수 $f_X(t) = F'_X(t)$	보렐측도 $(\mu \circ g^{-1})$ 의 라돈-니코딤미분 $\frac{d(\mu \circ g^{-1})}{d\lambda}(t) = F'_g(t)$
Law of the Unconscious Statistician $\mathbb{E}(h(X)) = \int h(t)f_X(t)dt$	합성함수 $(h \circ g)$ 의 르벡적분과 라돈-니코딤미분의 성질 $\int (h \circ g) d\mu = \int h(t) \frac{d(\mu \circ g^{-1})}{d\lambda}(t) dt$





여러분의 생각은 어느 쪽에 더 가깝나요?

1. Vitali의 집합  $V$ 처럼 사회성 떨어지는 예시를 물어둔다면, 기초적인 통계 지식으로도 수학과 4학년생들이 다루는 개념들을 **활용**할 수 있구나!



여러분의 생각은 어느 쪽에 더 가깝나요?

1. Vitali의 집합  $V$ 처럼 사회성 떨어지는 예시를 물어둔다면, 기초적인 통계 지식으로도 수학과 4학년생들이 다루는 개념들을 **활용**할 수 있구나!
  - 통계를 더 공부하면 (수학만 공부하는 경우보다는) 과학 전반에서 발견할 수 있는 여러 현상을 체계적으로 잘 분석할 수 있을 거야.



여러분의 생각은 어느 쪽에 더 가깝나요?

1. Vitali의 집합  $V$ 처럼 사회성 떨어지는 예시를 물어둔다면, 기초적인 통계 지식으로도 수학과 4학년생들이 다루는 개념들을 **활용**할 수 있구나!
  - 통계학을 더 공부하면 (수학만 공부하는 경우보다는) 과학 전반에서 발견할 수 있는 여러 현상을 체계적으로 잘 분석할 수 있을 거야.
2. 통계학과에 가면 졸업할 때까지 확률의 **엄밀한** 정의를 모를 수도 있구나!



여러분의 생각은 어느 쪽에 더 가깝나요?

1. Vitali의 집합  $V$ 처럼 사회성 떨어지는 예시를 물어둔다면, 기초적인 통계 지식으로도 수학과 4학년생들이 다루는 개념들을 **활용**할 수 있구나!
  - 통계학을 더 공부하면 (수학만 공부하는 경우보다는) 과학 전반에서 발견할 수 있는 여러 현상을 체계적으로 잘 분석할 수 있을 거야.
2. 통계학과에 가면 졸업할 때까지 확률의 **엄밀한** 정의를 모를 수도 있구나!
  - 나는 모든 개념을 엄밀하게 정의내리면서,  $V$ 와 같은 독특한 예시를 더 공부할 거야.



## 통계학과 수학의 비교

수학, 통계학 사이에서 전공을 고민 중이라면?

- 벌써부터 선택을 후회할까 두려워하지는 마세요.





## 통계학과 수학의 비교

수학, 통계학 사이에서 전공을 고민 중이라면?

- 벌써부터 선택을 후회할까 두려워하지는 마세요.
- 사람의 생각은 바뀌기 마련입니다.





## 통계학과 수학의 비교

수학, 통계학 사이에서 전공을 고민 중이라면?

- 벌써부터 선택을 후회할까 두려워하지는 마세요.
- 사람의 생각은 바뀌기 마련입니다.
  - 수학 학사 통계 박사, 통계 학사 수학 박사인 사례 매우 많음!



# 통계학의 전문적 정의



정의	상식적	전문적
통계학	데이터를 분석하는 수학 분야	...
데이터	주어진 것들	확률변수의 모임
확률변수	무작위적인 변수	(가측)함수 $\Omega \rightarrow \mathbb{R}$
가측함수	상식적인 함수	사건... 확률...

통계학의 전문적인 정의는?

1. 확률을 다루는 학문?



# 통계학의 전문적 정의



정의	상식적	전문적
통계학	데이터를 분석하는 수학 분야	...
데이터	주어진 것들	확률변수의 모임
확률변수	무작위적인 변수	(가측)함수 $\Omega \rightarrow \mathbb{R}$
가측함수	상식적인 함수	사건... 확률...

통계학의 전문적인 정의는?

1. 확률을 다루는 학문?
2. 차원의 축소(dimension reduction)?

# 통계학의 전문적 정의



정의	상식적	전문적
통계학	데이터를 분석하는 수학 분야	...
데이터	주어진 것들	확률변수의 모임
확률변수	무작위적인 변수	(가측)함수 $\Omega \rightarrow \mathbb{R}$
가측함수	상식적인 함수	사건... 확률...

통계학의 전문적인 정의는?

1. 확률을 다루는 학문?
2. 차원의 축소(dimension reduction)?
3. 근사(approximation)와 오차(error)의 학문?

# 통계학의 전문적 정의



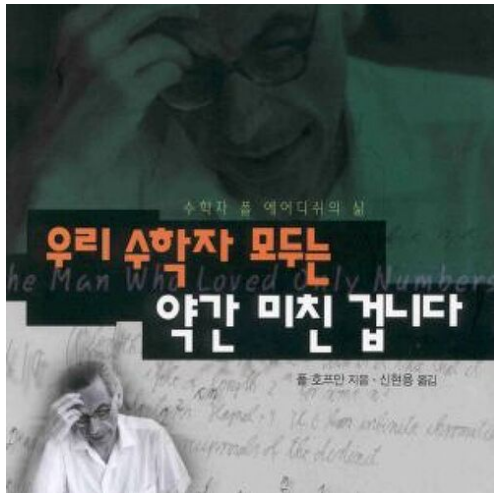
정의	상식적	전문적
통계학	데이터를 분석하는 수학 분야	...
데이터	주어진 것들	확률변수의 모임
확률변수	무작위적인 변수	(가측)함수 $\Omega \rightarrow \mathbb{R}$
가측함수	상식적인 함수	사건... 확률...

통계학의 전문적인 정의는?

1. 확률을 다루는 학문?
2. 차원의 축소(dimension reduction)?
3. 근사(approximation)와 오차(error)의 학문?
4. 의사 결정(decision making)의 수치화?

## 확률을 다루는 학문?

- “데이터 = 확률변수의 모임”이라는 시각에 집중
- 주어진 것들 = 함수??



# 확률을 다루는 학문?

- 확률을 엄밀하게 정의내리기는 어렵지만...



# 확률을 다루는 학문?



- 확률을 엄밀하게 정의내리기는 어렵지만...
- Henri Lebesgue(1875-1941), Andrey Kolmogorov(1903-1987)가 태어나기 한참 전부터 확률은 큰 관심을 받았음



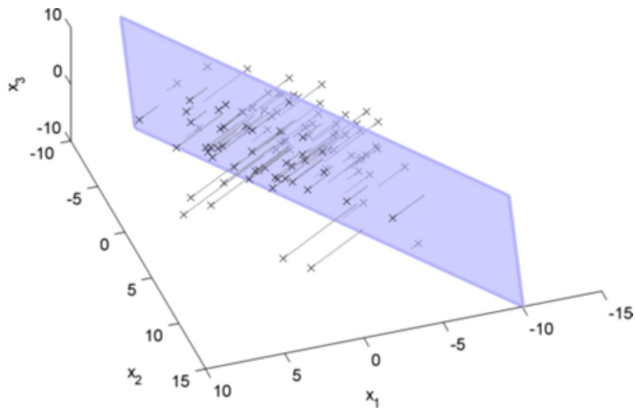
# 확률을 다루는 학문?



- 확률을 엄밀하게 정의내리기는 어렵지만...
- Henri Lebesgue(1875-1941), Andrey Kolmogorov(1903-1987)가 태어나기 한참 전부터 확률은 큰 관심을 받았음
- Girolamo Cardano(1501-1576)



# 차원의 축소?





# 차원의 축소?

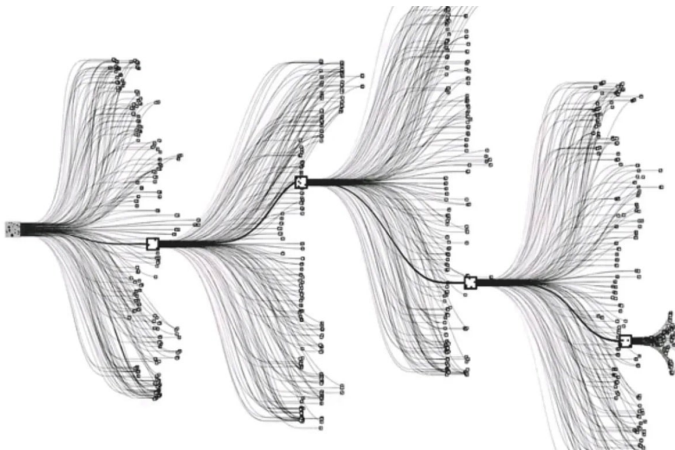


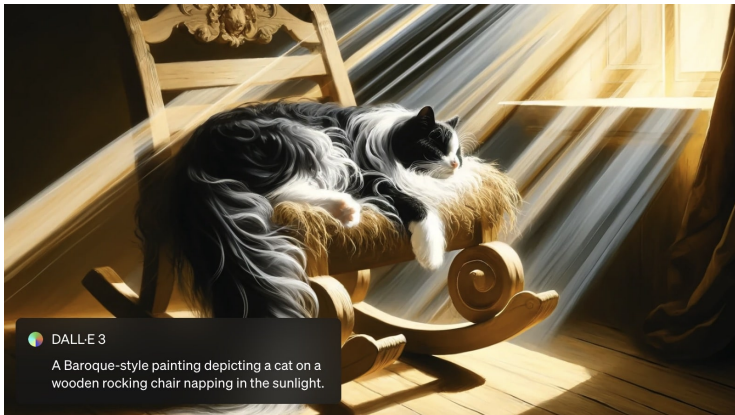
# 근사와 오차의 학문?



- 자원(돈, 시간, 전기 등)이 무한하다면 통계학은 필요가 없다.
- “이만하면 충분하오”

# 의사 결정의 수치화?

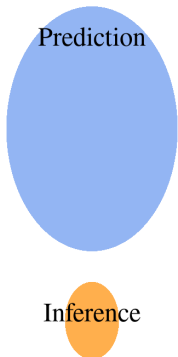




- 앞으로 무엇을 통계학에서 다뤄야 할까?



## Machine Learning

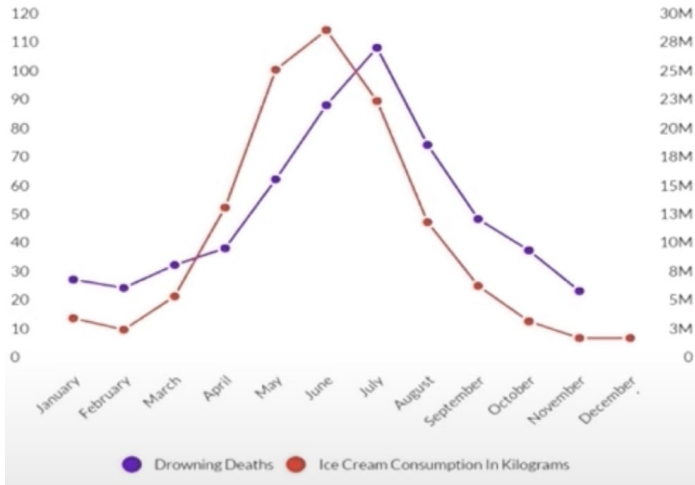


## Statistics





## 월별 익사 사망자 수와 아이스크림 소비량 (2018)





# 기타 중요한 통계학의 관심사



- 작은 데이터 학습
- 고차원 데이터 학습
- 생존 분석
- 결측치 처리
- 시계열 분석
- 로버스트 통계
- 실험 계획
- 모형 평가
- 정규성 검정
- 비모수 통계
- 개인정보 보호
- ...



“이만하면 충분해요”



Please visit <https://joonhyuk.com> for more information.



Contact: [joonhyukjung \(at\) uchicago.edu](mailto:joonhyukjung@uchicago.edu)

# Special Thanks



- 일부 사진은 서울대학교 통계학과 박건웅 교수님의 유튜브에서 가져왔습니다. 사용을 허락해주신 교수님께 감사드립니다.
- 백승용 선생님을 비롯하여 진로 탐색의 날을 마련해주신 서울과학고등학교 선생님들께 감사드립니다.
- 동창회 학술부장 24기 이재령 선배님께 감사드립니다.
- 마지막으로, 강연 들어주신 설곽인 모두에게 감사드립니다.