

기초 통계

for Manuscript Editor

2021년 11월 4일

박근철

1. 논문 내 통계
 - a. 화이자 백신 효과 95% ?
2. 설문조사로 본 신뢰구간
 - a. 2·4 부동산 대책 '도움되지 않을 것' $53.1\% \pm 4.4\%$ p (설문조사)
3. 논문작성에 필요한 기초통계 학습
 - a. 데이터
 - b. 대표값
 - c. p값
4. 논문작성에 사용되는 통계S/W 소개
5. ME자격시험 교육용 문제 풀이

화이자 백신 효과 95% ?

The screenshot shows the NEJM website interface. At the top, there's a navigation bar with the NEJM logo and a search bar. Below this, there's a section for 'PERSPECTIVE' with a link to 'Addressing Child Hunger When School Is Closed — Considerations during the Pandem...'. The main article is titled 'Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine' by Fernando P. Polack, M.D., et al. The article is dated December 31, 2020. The abstract states: 'Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection and the resulting coronavirus disease 2019 (Covid-19) have afflicted tens of millions of people in a worldwide pandemic. Safe and effective vaccines are needed urgently.' The methods section describes a multinational, placebo-controlled, observer-blinded, pivotal efficacy trial. The article is available in English, Chinese, and Spanish. The NEJM CareerCenter logo is visible at the bottom right.

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

Abstract : Background, Methods, Results, Conclusions

Introduction Methods

TRIAL OBJECTIVES
PARTICIPANTS AND OVERSIGHT
TRIAL PROCEDURES
SAFETY
EFFICACY
STATISTICAL ANALYSIS

Results

PARTICIPANTS
SAFETY
ADVERSE EVENTS(부작용)
EFFICACY

Discussion

Method : 연구방법

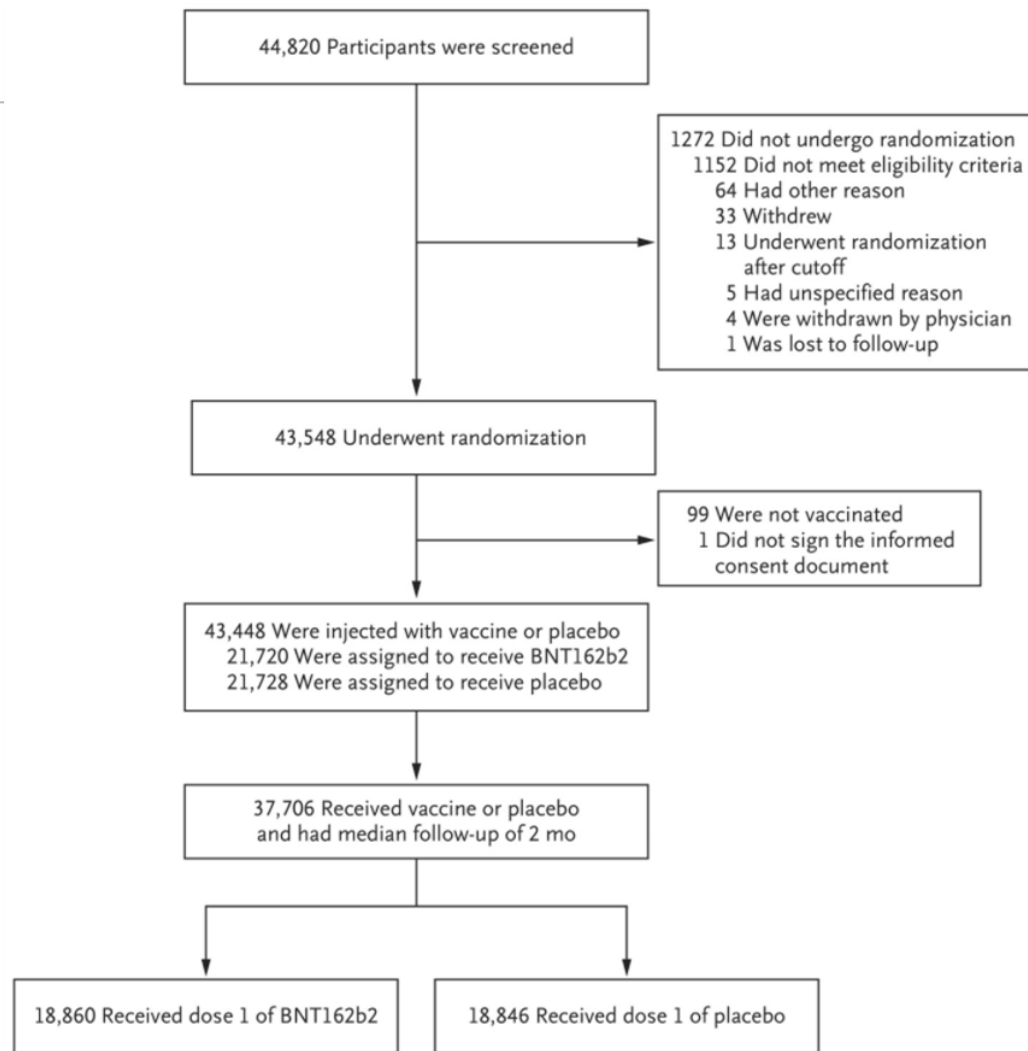
placebo-controlled, observer-blinded, pivotal efficacy trial, we **randomly assigned** persons 16 years of age or older in a 1:1 ratio to receive two doses, 21 days apart, of either placebo or the BNT162b2 vaccine candidate (30 µg per dose).

- placebo-controlled : 대조군, 실험군
전체 실험 참가자를 대조군과 실험군으로 나누어서, 실험군에는 백신후보물질을 접종하고, 대조군에는 위약(placebo)을 접종.
- observer-blind : 관찰자도 모름
누구에게 백신후보물질을 접종하고, 위약을 접종하는지 관찰자도 모르게 함.
- randomly assigned : 무작위배정
대조군과 실험군에 배정될 확률을 동일하게 해서 배정함.

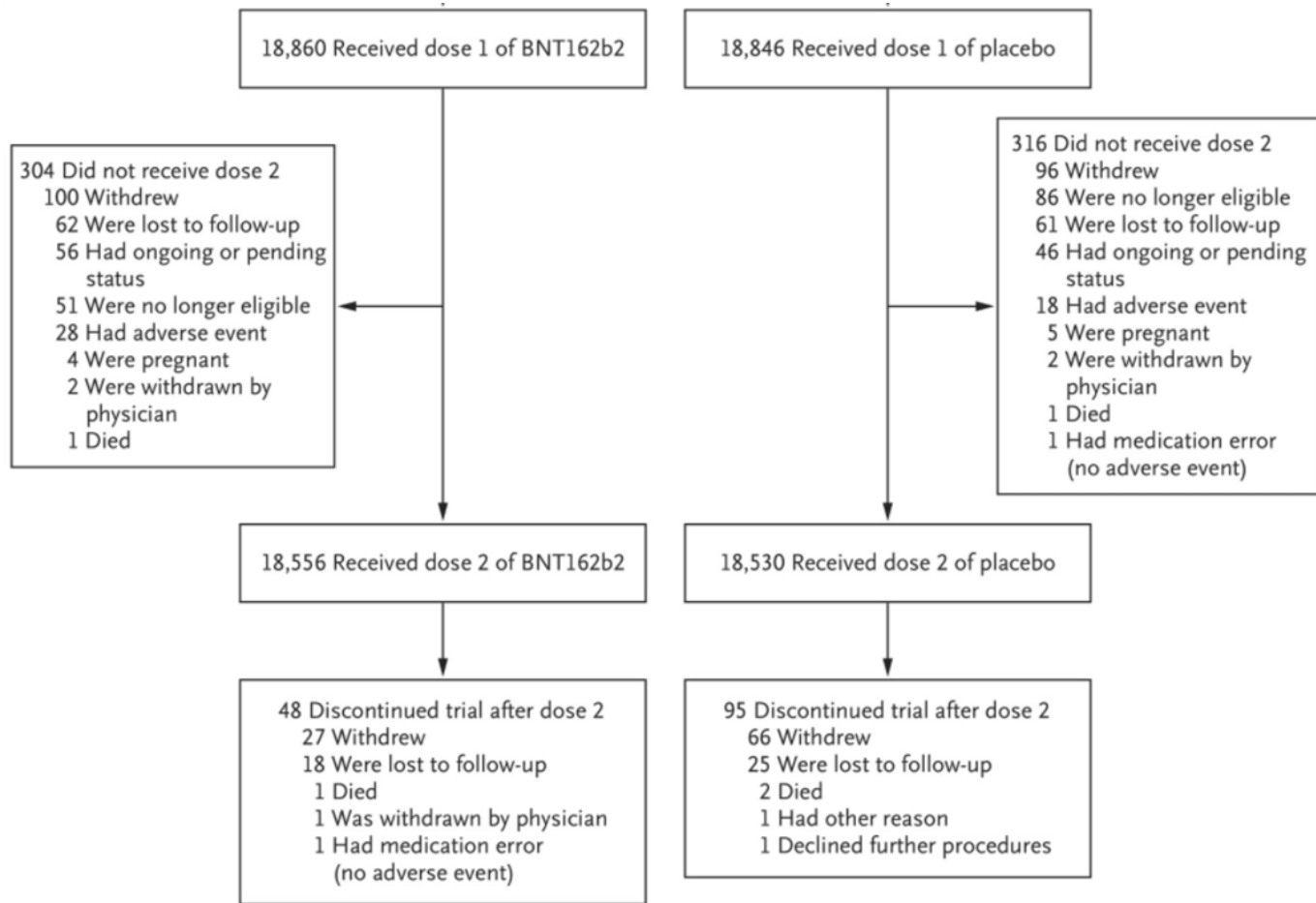
Results : 참가자*

Figure 1. Enrollment and Randomization.

The diagram represents all enrolled participants through November 14, 2020. The safety subset (those with a median of 2 months of follow-up, in accordance with application requirements for Emergency Use Authorization) is based on an October 9, 2020, data cut-off date. The further procedures that one participant in the placebo group declined after dose 2 (lower right corner of the diagram) were those involving collection of blood and nasal swab samples.



Results : 참가자(계속)*



Results : 참가자(계속)*

Table 1. Demographic Characteristics of the Participants in the Main Safety Population.

* Percentages may not total 100 because of rounding.

† Race or ethnic group was reported by the participants.

‡ The body-mass index is the weight in kilograms divided by the square of the height in meters.

몸무게(kg)를 키(m)의 제곱으로 나눔

Characteristic	BNT162b2(N=18,860)	Placebo(N=18,846)	Total(N=37,706)
Sex — no. (%)			
Male	9,639 (51.1)	9,436 (50.1)	19,075 (50.6)
Female	9,221 (48.9)	9,410 (49.9)	18,631 (49.4)
Race or ethnic group — no. (%)†			
White	15,636 (82.9)	15,630 (82.9)	31,266 (82.9)
Black or African American	1,729 (9.2)	1,763 (9.4)	3,492 (9.3)
Asian	801 (4.2)	807 (4.3)	1,608 (4.3)
Native American or Alaska Native	102 (0.5)	99 (0.5)	201 (0.5)
Native Hawaiian or other Pacific Islander	50 (0.3)	26 (0.1)	76 (0.2)
Multiracial	449 (2.4)	406 (2.2)	855 (2.3)
Not reported	93 (0.5)	115 (0.6)	208 (0.6)
Hispanic or Latinx	5,266 (27.9)	5,277 (28.0)	10,543 (28.0)
Country — no. (%)			
Argentina	2,883 (15.3)	2,881 (15.3)	5,764 (15.3)
Brazil	1,145 (6.1)	1,139 (6.0)	2,284 (6.1)
South Africa	372 (2.0)	372 (2.0)	744 (2.0)
United States	14,460 (76.7)	14,454 (76.7)	28,914 (76.7)
Age group — no. (%)			
16–55 yr	10,889 (57.7)	10,896 (57.8)	21,785 (57.8)
>55 yr	7,971 (42.3)	7,950 (42.2)	15,921 (42.2)
Age at vaccination — yr			
Median	52.0	52.0	52.0
Range	16–89	16–91	16–91
Body-mass index‡			
≥30.0: obese	6,556 (34.8)	6,662 (35.3)	13,218 (35.1)

실험결과 : Result

Table 2. Vaccine Efficacy against Covid-19 at Least 7 days after the Second Dose.

Efficacy End Point	BNT162b2		Placebo		Vaccine Efficacy, % (95% Credible Interval) [‡]	Posterior Probability (Vaccine Efficacy >30%) [§]
	No. of Cases	Surveillance Time (n) [†]	No. of Cases	Surveillance Time (n) [†]		
	(N=18,198)		(N=18,325)			
Covid-19 occurrence at least 7 days after the second dose in participants without evidence of infection	8	2.214 (17,411)	162	2.222 (17,511)	95.0 (90.3–97.6)	>0.9999
	(N=19,965)		(N=20,172)			
Covid-19 occurrence at least 7 days after the second dose in participants with and those without evidence of infection	9	2.332 (18,559)	169	2.345 (18,708)	94.6 (89.9–97.3)	>0.9999

* The total population without baseline infection was 36,523; total population including those with and those without prior evidence of infection was 40,137.

† The surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from 7 days after the second dose to the end of the surveillance period.

‡ The credible interval for vaccine efficacy was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

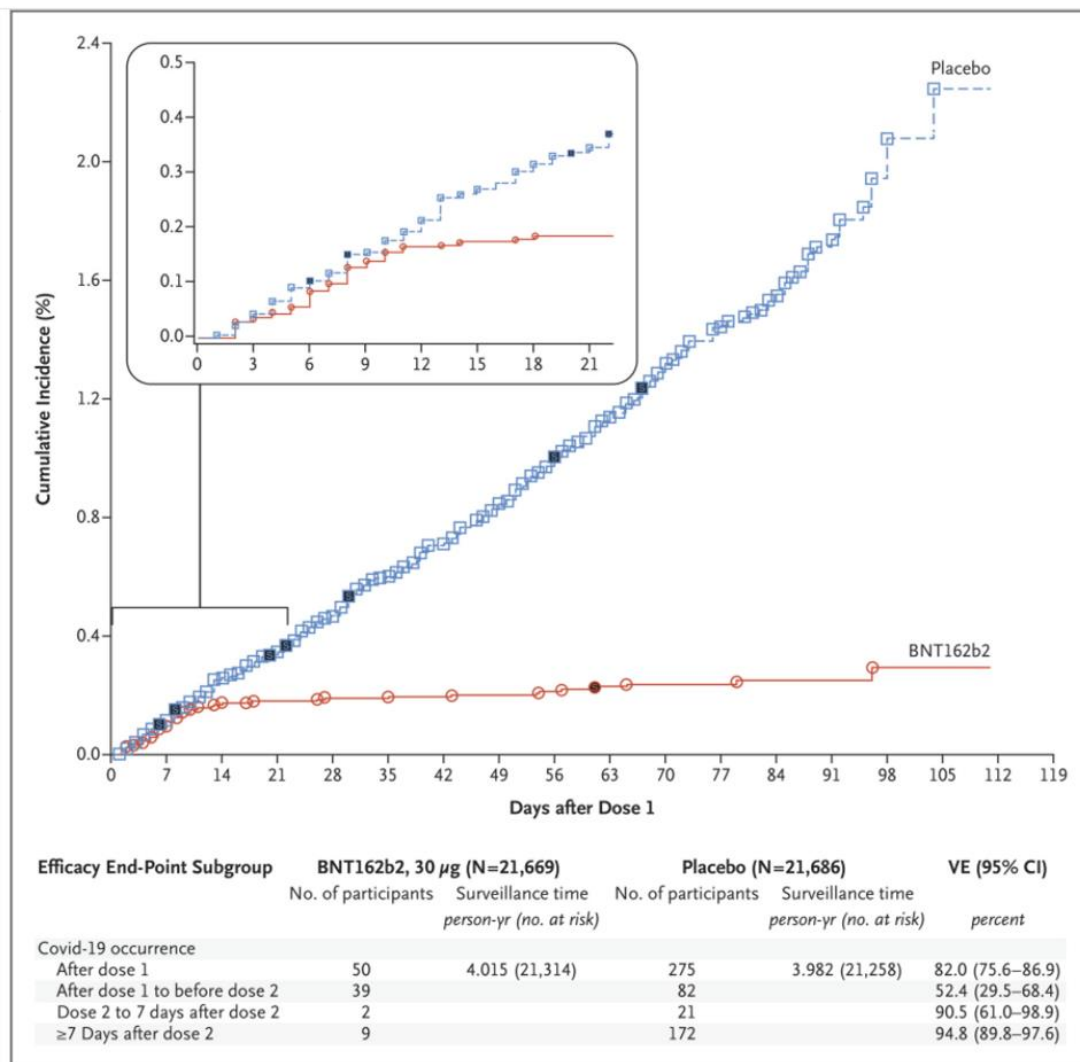
§ Posterior probability was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

실험결과 : Result

Figure 3. Efficacy of BNT162b2 against Covid-19 after the First Dose.

Shown is the cumulative incidence of Covid-19 after the first dose (modified intention-to-treat population). Each symbol represents Covid-19 cases starting on a given day; filled symbols represent severe Covid-19 cases. Some symbols represent more than one case, owing to overlapping dates. The inset shows the same data on an enlarged y axis, through 21 days. Surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from the first dose to the end of the surveillance period. The confidence interval (CI) for vaccine efficacy (VE) is derived according to the Clopper–Pearson method.

첫 번째 접종 후 Covid-19의 누적 발생률이 표시됩니다. 각 기호는 주어진 날짜에 시작되는 Covid-19 (감염) 사례를 나타냅니다.

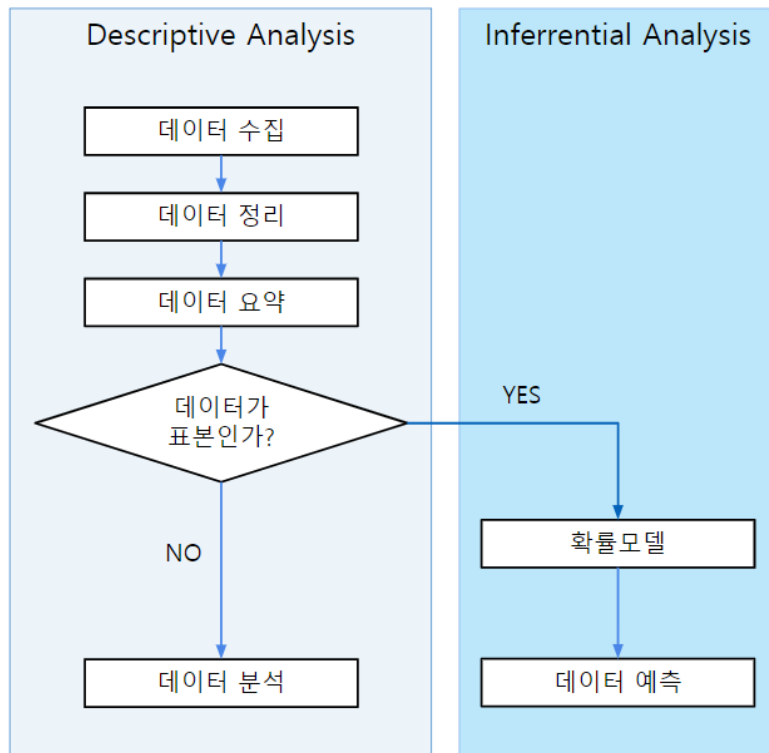


통계

기술통계(Descriptive Statistics)

추론통계(Inference Statistics) : 표본(Sample)으로 모집단(Population)을 추론.

임상시험에 참여한 사람들의 통계량(Statistics)으로 전체 인류에 대한 모수(Parameter)를 추론



신용구간(Credible Interval)

Table 2. Vaccine Efficacy against Covid-19 at Least 7 days after the Second Dose.

Efficacy End Point	BNT162b2		Placebo		Vaccine Efficacy, % (95% Credible Interval) [‡]	Posterior Probability (Vaccine Efficacy >30%) [§]
	No. of Cases	Surveillance Time (n) [‡]	No. of Cases	Surveillance Time (n) [‡]		
	(N=18,198)		(N=18,325)			
Covid-19 occurrence at least 7 days after the second dose in participants without evidence of infection	8	2.214 (17,411)	162	2.222 (17,511)	95.0 (90.3–97.6)	>0.9999
	(N=19,965)		(N=20,172)			
Covid-19 occurrence at least 7 days after the second dose in participants with and those without evidence of infection	9	2.332 (18,559)	169	2.345 (18,708)	94.6 (89.9–97.3)	>0.9999

‡ The credible interval for vaccine efficacy was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

(90.3–97.6)

2.4 부동산 대책 '도움되지 않을 것'
53.1%±4.4%p

설문조사

지난 2월 4일 정부가 2025년까지 83만 6천호를 짓기 위한 주택 부지 공급안을 발표했는데, 이 공급안에 대한 여론조사 전문기관의 조사 결과 “도움이 되지 않을 것이다”라는 응답이 53.1%, “도움이 될 것이다”라는 응답은 41.7%였습니다. 이 결과는 2월 5일 여론조사 전문기관에서 접촉한 전국의 만18세 이상 6,735명으로부터 응답한 최종 500명의 답변을 요약한 것입니다.

여론조사 전문기관에서는 2·4 부동산 대책, ‘도움 되지 않을 것’ 53.1%, 표본오차는 95% 신뢰수준에서 $\pm 4.4\%$ p라고 발표했습니다.

	응답자 수
도움이 될 것이다.	208.5(41.7%)
도움이 되지 않을 것이다.	265.5(53.1%)
잘 모르겠다	26(5.2%)
합계	500(100.0%)

그럼, 우리나라 만18세 이상 전체 국민 중 몇 %가 ‘도움되지 않을 것’이라고 판단할까요?

설문조사 신뢰구간(Confidence Interval)

지난 2월 4일 정부가 2025년까지 83만 6천호를 짓기 위한 주택 부지 공급안을 발표했는데, 이 공급안에 대한 여론조사 전문기관의 조사 결과 “도움이 되지 않을 것이다”라는 응답이 53.1%, “도움이 될 것이다”라는 응답은 41.7%였습니다. 이 결과는 2월 5일 여론조사 전문기관에서 접촉한 전국의 만18세 이상 6,735명으로부터 응답한 최종 500명의 답변을 요약한 것입니다.

여론조사 전문기관에서는 2·4 부동산 대책, ‘도움 되지 않을 것’ 53.1%, 표본오차는 95% 신뢰수준에서 ±4.4%p라고 발표했습니다.

	응답자 수
도움이 될 것이다.	208.5(41.7%)
도움이 되지 않을 것이다.	265.5(53.1%)
잘 모르겠다	26(5.2%)
합계	500(100.0%)

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

$$\left[\hat{p} - 4.4\%p, \hat{p} + 4.4\%p \right]$$

$$\hat{p} = 0.531 (53.1\%)$$

$$z_{\alpha/2} = 1.96 \dots (95\% \text{ 신뢰수준의 임계값})$$

$$n = 500$$

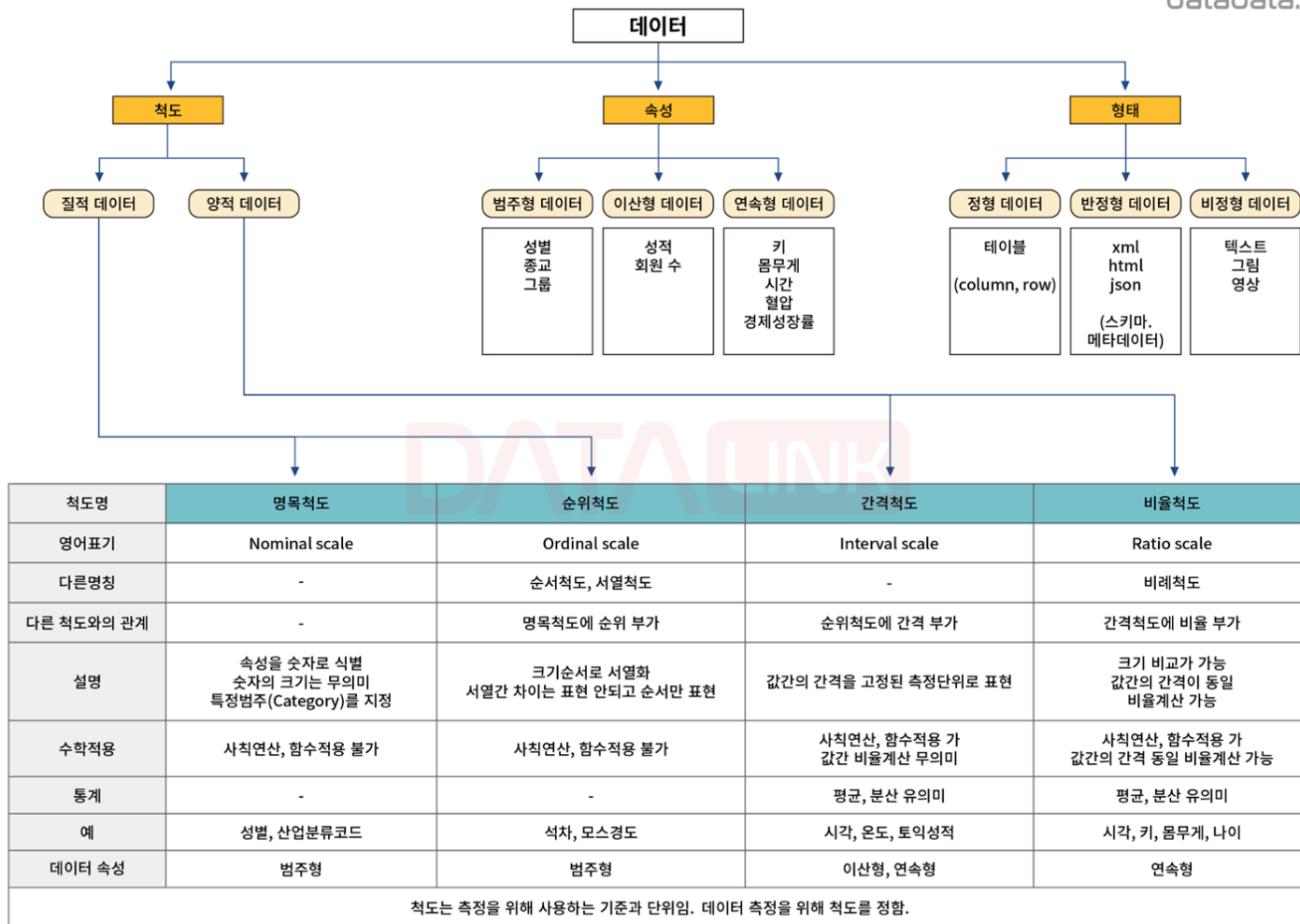
$$(48.7 - 57.5)$$

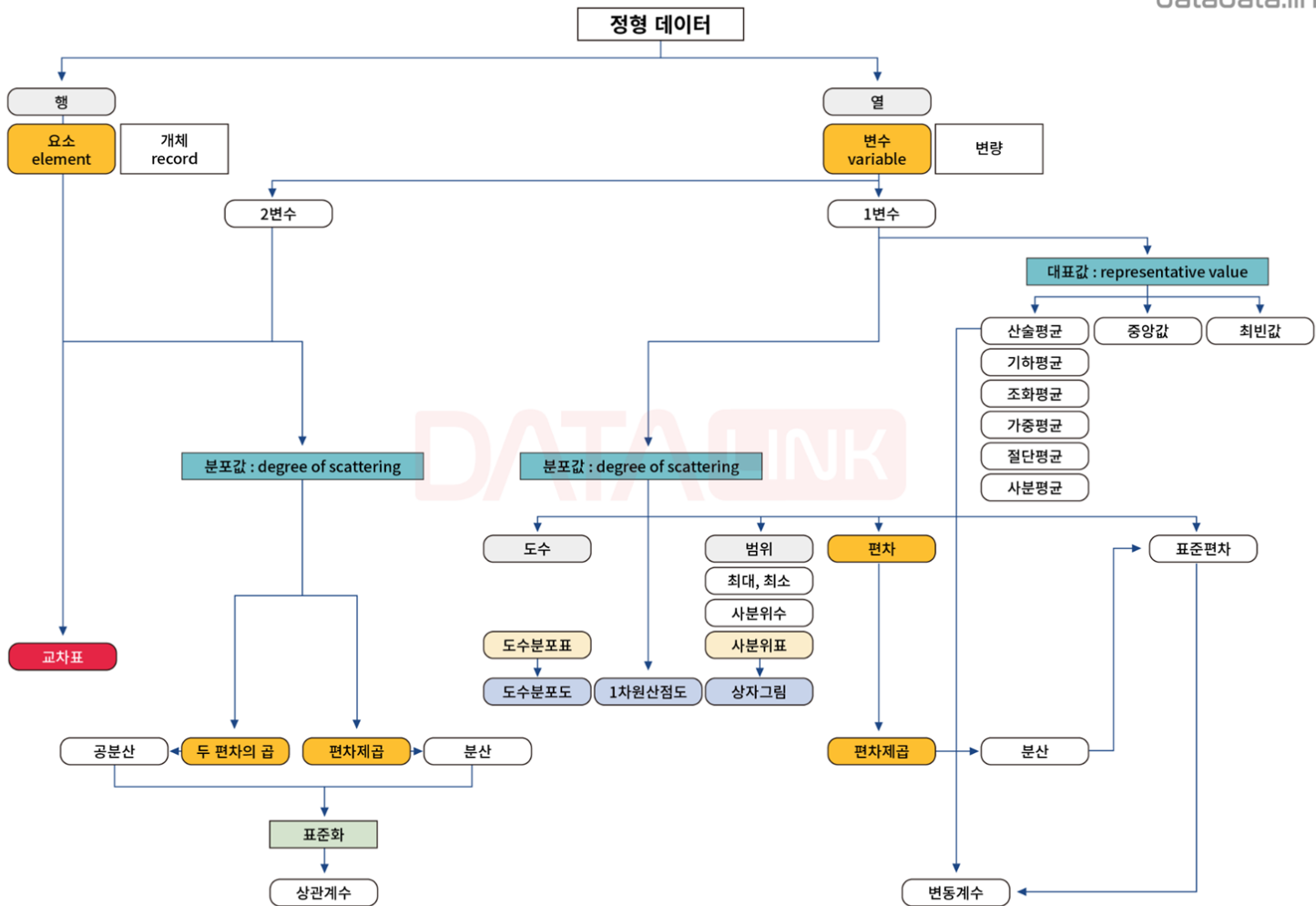
설문조사 신뢰구간(Confidence Interval)

우리나라 만 18세 이상 국민들 중 48.7~57.5%는
2.4 부동산 대책 '도움되지 않을 것'이라고 판단할 것입니다.
(95% 신뢰구간. n=500명)

기초통계*

데이터종류





데이터의 대표값 예제 1

예제 질문

한국과학학술지편집인협의회 ME의 연봉은 얼마인가요?
하나의 숫자로 답변해주세요.

예제 데이터

ME 5명의 연봉 데이터 : 1, 2, 3, 4, 10(단위 : 천만원)

답변

합계 : 20(천만원)

개수 : 5

평균 : 4(천만원)

중앙값 : 3(천만원)

데이터의 대표값

평균(mean)

데이터의 합계를 데이터의 개수로 나눈 값 : 산술평균

$$\text{평균} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

중앙값(median)

데이터를 크기 순서로 나열할 때 중앙에 놓이는 값

$$\text{중앙값} = \left\{ \frac{(n+1)}{2} \text{ 번째 데이터, } n \text{ 이 홀수인 경우} \right.$$

$$\text{중앙값} = \left\{ \left(\frac{n}{2} \right) \text{ 번째와 } \left(\frac{n+1}{2} \right) \text{ 번째 데이터의 평균, } n \text{ 이 짝수인 경우} \right.$$

최빈값(mode)

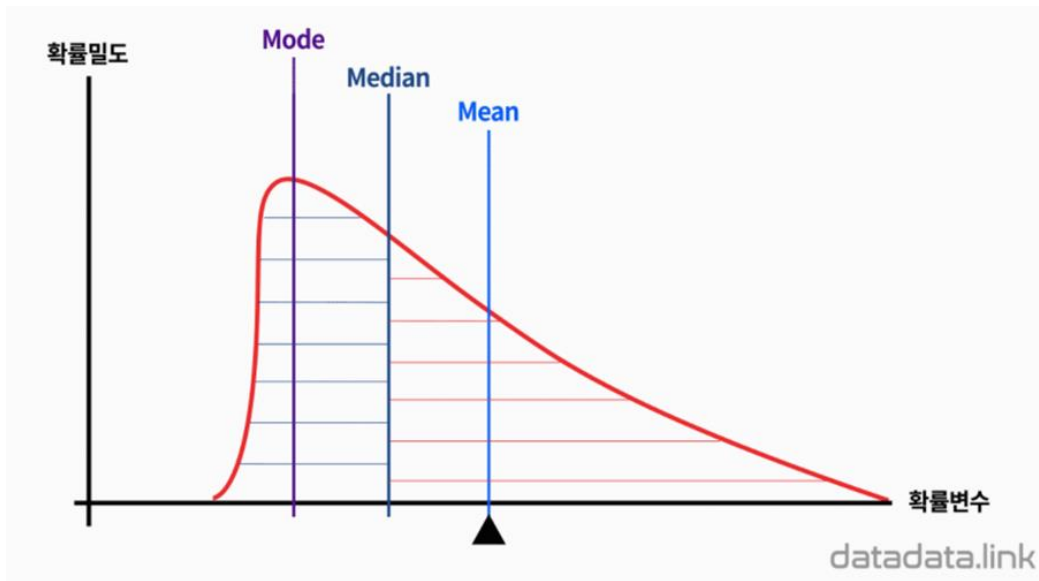
데이터 중 가장 빈도가 많은 값

데이터의 대표값

국회의원 재산 평균 94억원?

2013년 3월 29일, 국회 공직자윤리위원회가 공개한 296명의 국회의원 재산 평균(산술평균)은 94억 9000만원입니다.

그런데, 정몽준 의원, 고희선 의원을 제외하고, 평균을 계산하면 23억 3000만원이었습니다. 두 의원의 재산은 각각 1조 9249만원, 1984억원이었습니다.



데이터의 대표값 예제 2

예제 질문

한국과학학술지편집인협의회 ME의 연봉은 얼마인가요?
두개의 숫자로 답변해주세요.

예제 데이터

ME 5명의 연봉 데이터 : 1, 2, 3, 4, 10(단위 : 천만원)

답변

합계 : 20(천만원)

개수 : 5

범위 : 1~10(천만원)

평균과 표준편차 : 4(천만원) ± 3

데이터의 대표값

범위(range)

데이터의 최대값과 최소값으로 표현. 혹은 최대값과 최소값의 차이로 표현.

분산(variance)

각 데이터와 평균과의 차이를 제곱해서 모두 더한 후, 데이터 개수 혹은 자유도로 나눈 값

분산(variance)

$$\text{모분산 } \sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad (N: \text{모집단 데이터수})$$

$$\text{표본분산 } s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (n: \text{표본 데이터수})$$

표준편차(standard deviation)

분산의 제곱근

표준편차(standard deviation)

$$\text{모표준편차 } \sigma = \sqrt{\sigma^2} \quad \text{표본표준편차 } s = \sqrt{s^2} \quad \text{분산의 제곱근}$$

사분위수범위(interquartile range)

데이터를 높은 값에서 작은 순서로 배열했을 때, 전체를 4등분하는 값을 사분위수라고 함. 사분위수 범위는 25%, 75%에 해당하는 데이터의 범위 혹은 그 차이로 표시.

Results : 참가자(계속)*

Table 1. Demographic Characteristics of the Participants in the Main Safety Population.

* Percentages may not total 100 because of rounding.

† Race or ethnic group was reported by the participants.

‡ The body-mass index is the weight in kilograms divided by the square of the height in meters.

몸무게(kg)를 키(m)의 제곱으로 나눔

Characteristic	BNT162b2(N=18,860)	Placebo(N=18,846)	Total(N=37,706)
Sex — no. (%)			
Male	9,639 (51.1)	9,436 (50.1)	19,075 (50.6)
Female	9,221 (48.9)	9,410 (49.9)	18,631 (49.4)
Race or ethnic group — no. (%)†			
White	15,636 (82.9)	15,630 (82.9)	31,266 (82.9)
Black or African American	1,729 (9.2)	1,763 (9.4)	3,492 (9.3)
Asian	801 (4.2)	807 (4.3)	1,608 (4.3)
Native American or Alaska Native	102 (0.5)	99 (0.5)	201 (0.5)
Native Hawaiian or other Pacific Islander	50 (0.3)	26 (0.1)	76 (0.2)
Multiracial	449 (2.4)	406 (2.2)	855 (2.3)
Not reported	93 (0.5)	115 (0.6)	208 (0.6)
Hispanic or Latinx	5,266 (27.9)	5,277 (28.0)	10,543 (28.0)
Country — no. (%)			
Argentina	2,883 (15.3)	2,881 (15.3)	5,764 (15.3)
Brazil	1,145 (6.1)	1,139 (6.0)	2,284 (6.1)
South Africa	372 (2.0)	372 (2.0)	744 (2.0)
United States	14,460 (76.7)	14,454 (76.7)	28,914 (76.7)
Age group — no. (%)			
16–55 yr	10,889 (57.7)	10,896 (57.8)	21,785 (57.8)
>55 yr	7,971 (42.3)	7,950 (42.2)	15,921 (42.2)
Age at vaccination — yr			
Median	52.0	52.0	52.0
Range	16–89	16–91	16–91
Body-mass index‡			
≥30.0: obese	6,556 (34.8)	6,662 (35.3)	13,218 (35.1)

p값 ?

나는 참이슬과 처음처럼의 맛을 구별할 수 있다.

사례(계속)

나는 참이슬과 처음처럼의 맛을 구별할 수 있다.

증명하고자 하는 것과 반대되는 가설을 세운다

참이슬과 처음처럼의 맛을 구별할 수 없다.

귀무가설, 영가설(Null Hypothesis, H_0)

귀무가설이 옳다는 가정 하에 대립되는 사건이 일어날 확률을 측정

참이슬과 처음처럼의 맛을 구별할 수 없는데도 불구하고, 어느 수준 이상 구별해 낸다면, 귀무가설이 틀린 것이 아닐까?

실험방식 1

사건	확률	신뢰수준
1회 구분	50.00%	50.00%
2회 연속으로 구분	25.00%	75.00%
3회 연속으로 구분	12.50%	87.50%
4회 연속으로 구분	6.25%	93.75%
5회 연속으로 구분	3.13%	96.88%

p값
(p value)

유의 확률(有意確率,
significance probability)

실험방식 2

사건	확률	신뢰수준
10번 중 0번 이상 구분	100.00%	0.00%
10번 중 1번 이상 구분	99.90%	0.10%
10번 중 2번 이상 구분	98.93%	1.07%
10번 중 3번 이상 구분	94.53%	5.47%
10번 중 4번 이상 구분	82.81%	17.19%
10번 중 5번 이상 구분	62.30%	37.70%
10번 중 6번 이상 구분	37.70%	62.30%
10번 중 7번 이상 구분	17.19%	82.81%
10번 중 8번 이상 구분	5.47%	94.53%
10번 중 9번 이상 구분	1.07%	98.93%
10번 중 10번 이상 구분	0.10%	99.90%

참이슬과 처음처럼의 맛을 구별할 수 없다.

귀무가설, 영가설(Null Hypothesis, H_0)

p값이 일정 수준 이하라면
귀무가설을 기각

참이슬과 처음처럼의 맛을 구별할 수 있다.

대립가설, 연구가설(Alternative Hypothesis, H_1)

다른 예

백신과 위약은 다르다.

증명하고자 하는 것과 반대되는 가설을 세운다

백신과 위약은 효능이 같다.
백신접종 or 위약접종이 COVID19에 걸리는 숫자의 차이는 0이다.
귀무가설, 영가설(Null Hypothesis, H_0)

귀무가설이 옳다는 가정하에 대립되는 사건이 일어날 확률을 측정

백신과 위약이 같은데도 불구하고, 어느 수준 이상 다르다면,
귀무가설이 틀린 것이 아닐까?

다른 예

남녀간 연봉이 다르다.

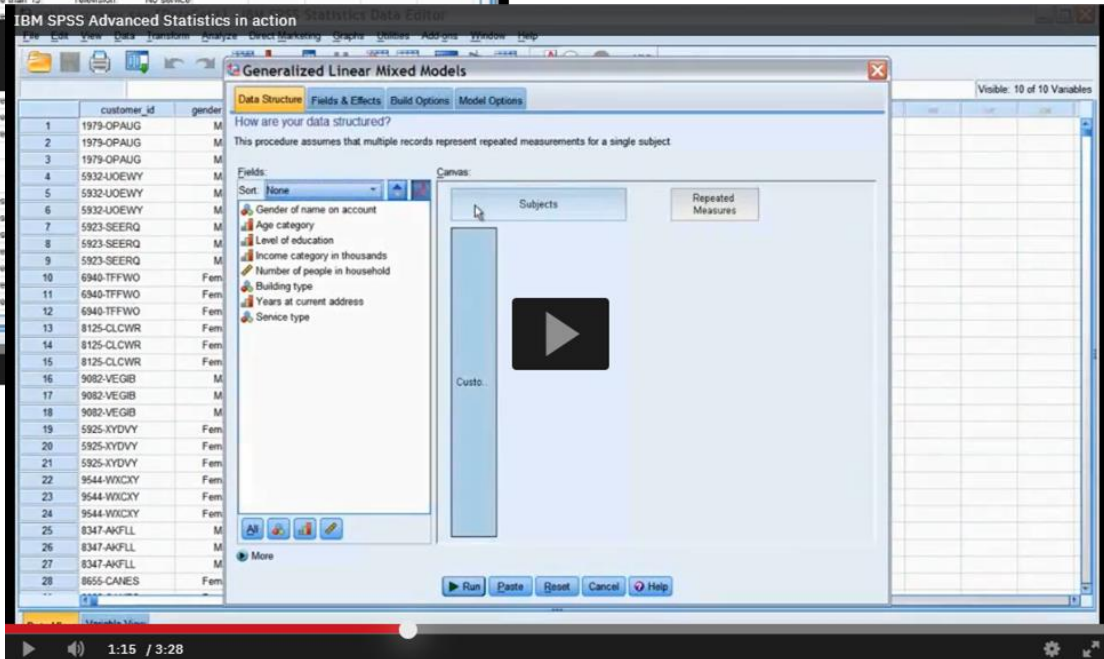
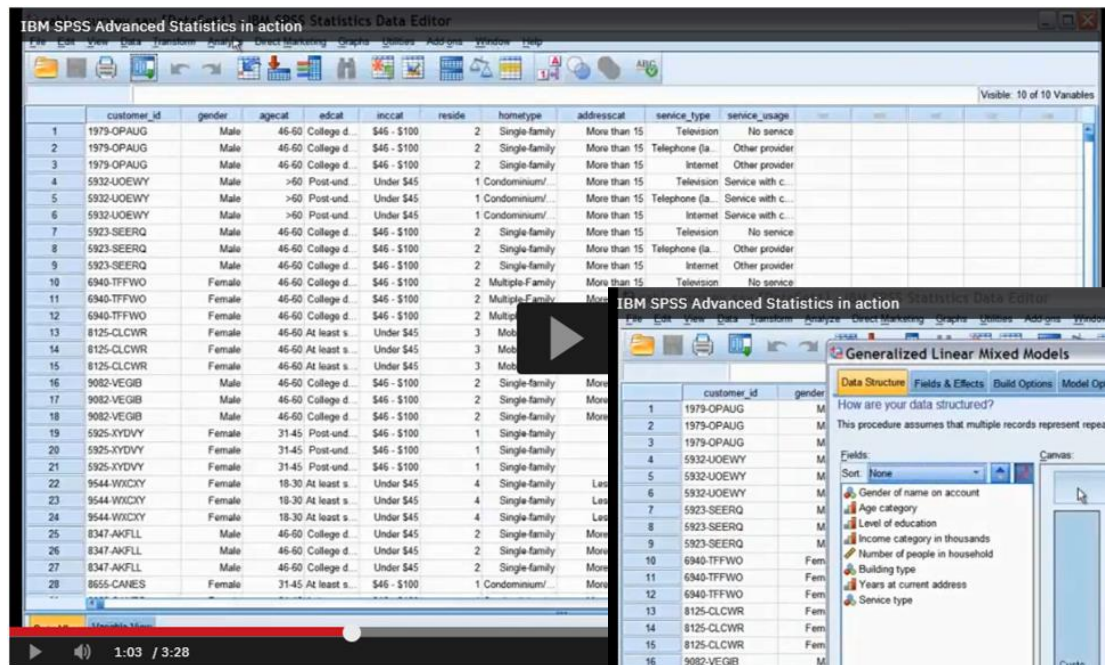
증명하고자 하는 것과 반대되는 가설을 세운다

남녀간 연봉이 같다.
남녀간 연봉의 차이는 0이다.
귀무가설, 영가설(Null Hypothesis, H_0)

귀무가설이 옳다는 가정하에 대립되는 사건이 일어날 확률을 측정

남녀간 연봉이 같은데도 불구하고, 어느 수준 이상 다르다면, 귀무가설이 틀린 것이 아닐까?

통계 소프트웨어 소개



R, R Studio

```
R Console

R version 4.0.0 (2020-04-24) -- "Arbor Day"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

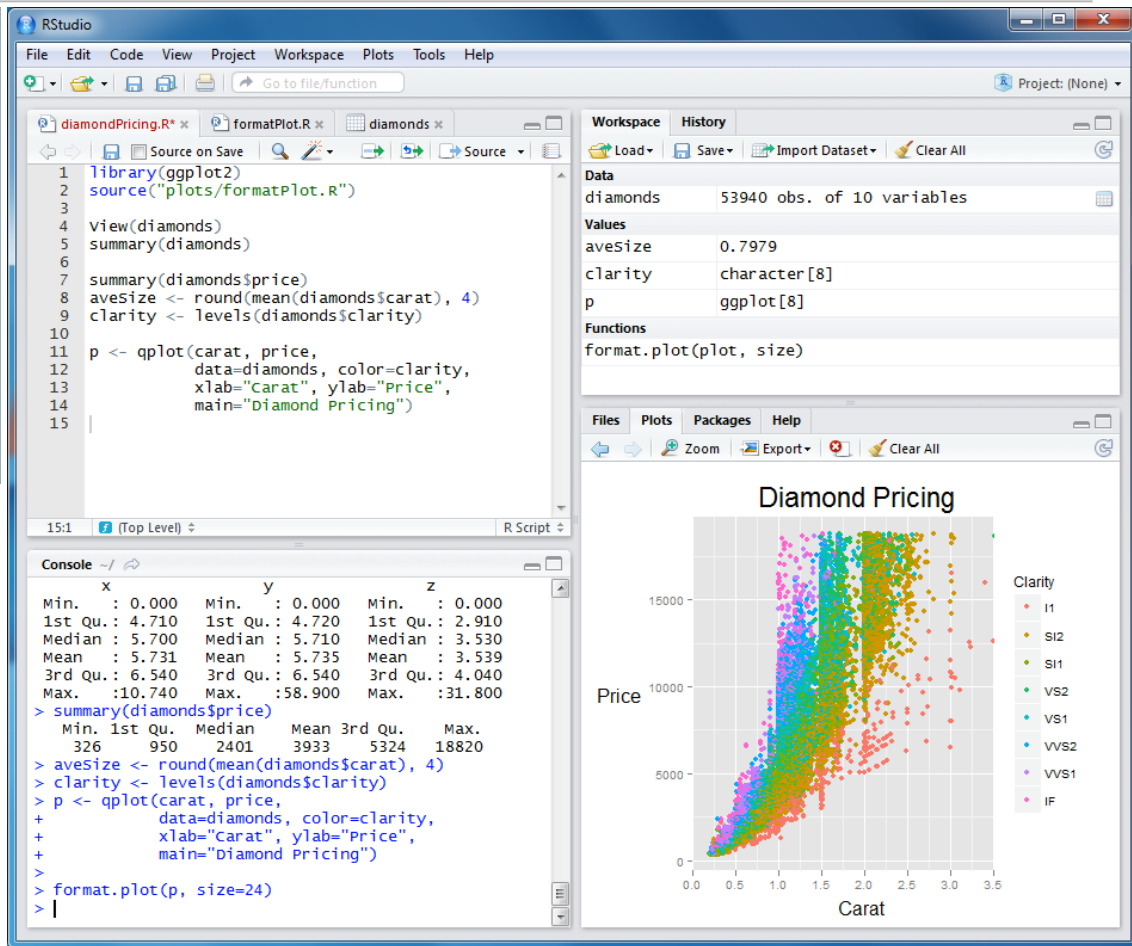
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

출처 1 : R
<https://www.r-project.org/>

출처 2 : RStudio IDE Features
<https://rstudio.com/products/rstudio/features/>



- SPSS
- R
- SAS
- Matlab
- Stata

SPSS(SPSS version 18.0, SPSS, Chicago, IL, USA)

SPSS(SPSS version 28.0, IBM, Armonk, NY, USA)

연습문제

1. Which of the following types of data involves only 2 categories?

- a. Continuous
- b. Discrete
- c. Ordinal
- d. Dichotomous
- e. Nominal

2. Which of the following types of data involves whole-number counts?

- a. Continuous
- b. Discrete
- c. Ordinal
- d. Dichotomous
- e. Nominal

3. Which of the following types of data includes a full range of possible fractional values?

- a. Continuous
- b. Discrete
- c. Ordinal
- d. Dichotomous
- e. Nominal

4. In a study of the association between physical activity and mental health, women were asked how many children they had. Which of the following levels of measurement best describes the variable number of children?

- a. Continuous
- b. Discrete
- c. Ordinal
- d. Dichotomous
- e. Nominal

5. Which of the following is the middle value (the 50th percentile) in a set of data?

- a. Mean
- b. Median
- c. Mode
- d. Range
- e. Standard deviation
- f. Interquartile range

6. Which of the following identifies the lowest and highest values in a set of data?

- a. Mean
- b. Median
- c. Mode
- d. Range
- e. Standard deviation
- f. Interquartile range

7. Which of the following identifies the 25th and 75th percentiles in a set of data?

- a. Mean
- b. Median
- c. Mode
- d. Range
- e. Standard deviation
- f. Interquartile range

8. Which of the following is the most frequent value in a set of data?

- a. Mean
- b. Median
- c. Mode
- d. Range
- e. Standard deviation
- f. Interquartile range

9. Which of the following is the average distance from the mean value in a set of data?

- a. Mean
- b. Median
- c. Mode
- d. Range
- e. Standard deviation
- f. Interquartile range

10. Which of the following terms means the probability of obtaining the observed data if the null hypothesis were exactly true?

- a. P value
- b. relative risk
- c. reliability
- d. true-positive rate

11. 다음의 표에서 ?에 들어갈 값을 구하세요. ?1부터 ?6까지 콤마로 분리해서 기입해주세요.

Table 1. Characteristics of the Participants

Category	Experiment Group	Control Group
Gender - no. (%)		
Male	52 (?1%)	?2 (?3%)
Female	?4 (?5%)	51 (?6%)
Total	100	100

연습문제해설

1. Which of the following types of data involves only **2 categories**?

- a. Continuous : 연속형
- b. Discrete : 이산형
- c. Ordinal : 순위
- d. Dichotomous : 이분형
- e. Nominal : 명목

예, 성별

2. Which of the following types of data involves **whole-number**(0과 자연수) **counts**?

- a. Continuous : 연속형
- b. Discrete : 이산형
- c. Ordinal : 순위형
- d. Dichotomous : 이분형
- e. Nominal : 명목형

natural number : 자연수, integer : 정수, fraction : 분수

3. Which of the following types of data includes a **full range of possible fractional values**(분수, 실수)?

- a. Continuous : 연속형
- b. Discrete : 이산형
- c. Ordinal : 순위형
- d. Dichotomous : 이분형
- e. Nominal : 명목형

natural number : 자연수, integer : 정수, fraction : 분수
예, 키, 몸무게

4. In a study of the association between physical activity and mental health, women were asked how many children they had. Which of the following levels of measurement best describes the variable **number of children**?

- a. Continuous : 연속형
- b. Discrete : 이산형
- c. Ordinal : 순위형
- d. Dichotomous : 이분형
- e. Nominal : 명목형

5. Which of the following is the **middle value (the 50th percentile)** in a set of data?

- a. Mean : 평균
- b. Median : 중앙값
- c. Mode : 최빈값
- d. Range : 범위(최대값과 최소값)
- e. Standard deviation : 표준편차
- f. Interquartile range : 사분위수 범위

6. Which of the following identifies the **lowest and highest** values in a set of data?

- a. Mean : 평균
- b. Median : 중앙값
- c. Mode : 최빈값
- d. Range : 범위(최대값과 최소값)
- e. Standard deviation : 표준편차
- f. Interquartile range : 사분위수 범위

7. Which of the following identifies the **25th and 75th percentiles** in a set of data?

- a. Mean : 평균
- b. Median : 중앙값
- c. Mode : 최빈값
- d. Range : 범위(최대값과 최소값)
- e. Standard deviation : 표준편차
- f. Interquartile range : 사분위수 범위

8. Which of the following is the **most frequent value** in a set of data?

- a. Mean : 평균
- b. Median : 중앙값
- c. Mode : 최빈값
- d. Range : 범위(최대값과 최소값)
- e. Standard deviation : 표준편차
- f. Interquartile range : 사분위수 범위

9. Which of the following is the **average distance from the mean** value in a set of data?

- a. Mean : 평균
- b. Median : 중앙값
- c. Mode : 최빈값
- d. Range : 범위(최대값과 최소값)
- e. Standard deviation : 표준편차
- f. Interquartile range : 사분위수 범위

10. Which of the following terms means the **probability** of obtaining the observed data **if the null hypothesis were exactly true?**

- a. **P value**
- b. relative risk
- c. reliability
- d. true-positive rate

11. 다음의 표에서 ?에 들어갈 값을 구하세요.

Table 1. Characteristics of the Participants

Category	Experiment Group	Control Group
Gender - no. (%)		
Male	52 (52%)	49 (49%)
Female	48 (48%)	51 (51%)
Total	100	100

심화

맥주는 손맛?



맥주는 손맛?



1. 파인트 잔을 깨끗하게 씻어 자연건조. 파인트잔의 온도는 맥주온도와 마찬가지로 5~8도.
2. 파인트 잔을 맥주가 나오는 꼭지에 45도로 기울이는 것은 흘러나오는 맥주의 거품이 너무 많이 생기지 않고 이상적인 대류(surging)를 일으키기 위해서이다. 이때 잔에 꼭지가 닿으면 안된다. 보통 맥주를 따르면 맥주의 거품이 위로 올라가지만 기네스의 경우 아래로 하강하는 것처럼 보인다. 드래프트 기네스에는 질소가스와 탄산가스가 7대3으로 혼합되어 있다. 이 혼합가스가 생맥주 꼭지에서 나오는 맥주와 섞여 대류를 만든다. 이때 질소의 거품입자가 탄산가스의 거품입자보다 작기 때문에 맥주를 아래로 밀어내면서 특유의 대류를 만든다.
3. 생맥주 탭의 손잡이를 몸 앞으로 90도 꺾어 파인트 잔의 3cm정도를 남기고 맥주를 한번에 따른다.
4. 맥주를 따른 후 119.5초동안 가스의 대류분리(surging)가 일어난다. 거품이 춤을 추듯이 솟아올라오는 '작은 폭포 쇼'를 보면서 기다리는 것도 기네스 맥주를 마시는 즐거움의 하나!
5. 119.5초가 지난 후 잔을 수직으로 세워 조심스럽게 손잡이를 뒤로 밀어 맥주를 채운다. 거품이 파인트 잔의 위에서 2mm정도까지 올라갈 때 맥주 따르기를 끝낸다.
6. 조금 기다리면 카푸치노의 크림과 같은 하얀 거품이 2cm정도 높아지고 맥주의 색깔은 서서히 진한 커피색으로 변한다. 이 때 맥주를 마신다.

맥주는 통계! : Student's t test

아일랜드의 기네스 맥주 회사에 근무하던 윌리엄 고셋은 **샘플(표본) 크기가 작은 샘플로 흠의 이상적인 비율을 연구**하던 중 논문을 발표

그의 필명을 따서 Student's t-distribution

- 필명을 사용해서 기네스 직원이라는 사실을 알리지 않음.
- 경쟁사에 기네스에서 사용하는 모델링을 알리고 싶지 않음.

The t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis.

귀무가설 하에서 스튜던트 t 분포를 따르는 통계량에 대한 검정

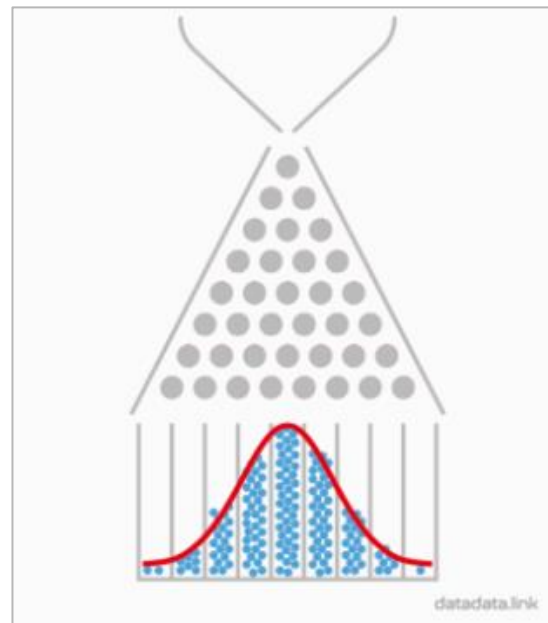
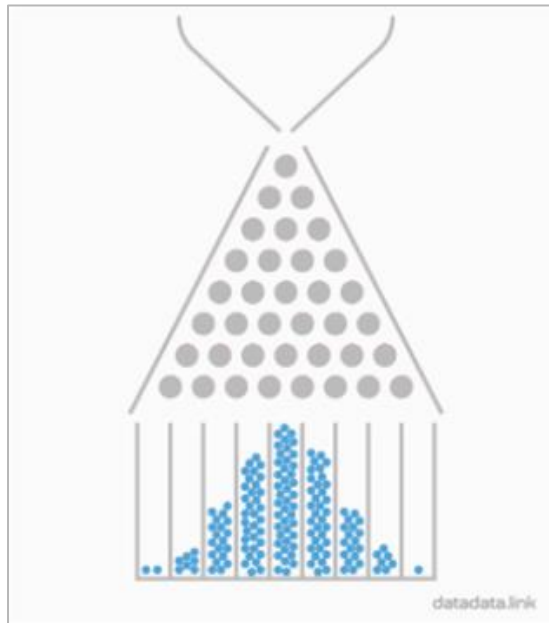
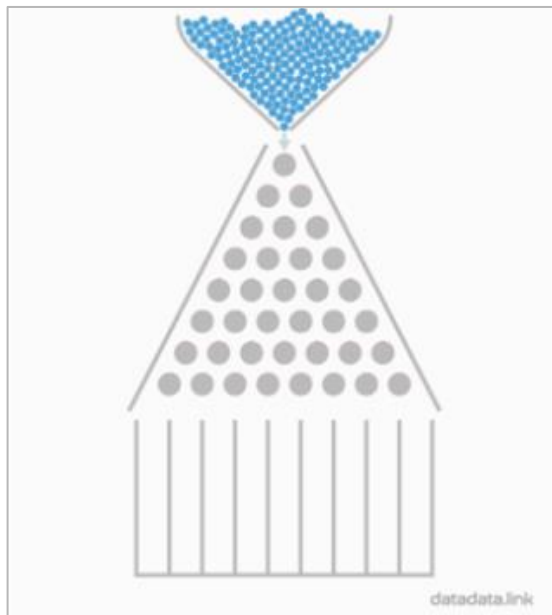
"Student" [William Sealy Gosset](#) (1908). "[The probable error of a mean](#)" (PDF). *Biometrika*. **6** (1): 1–25.
[doi:10.1093/biomet/6.1.1](#). [hdl:10338.dmlcz/143545](#). [JSTOR](#) 2331554

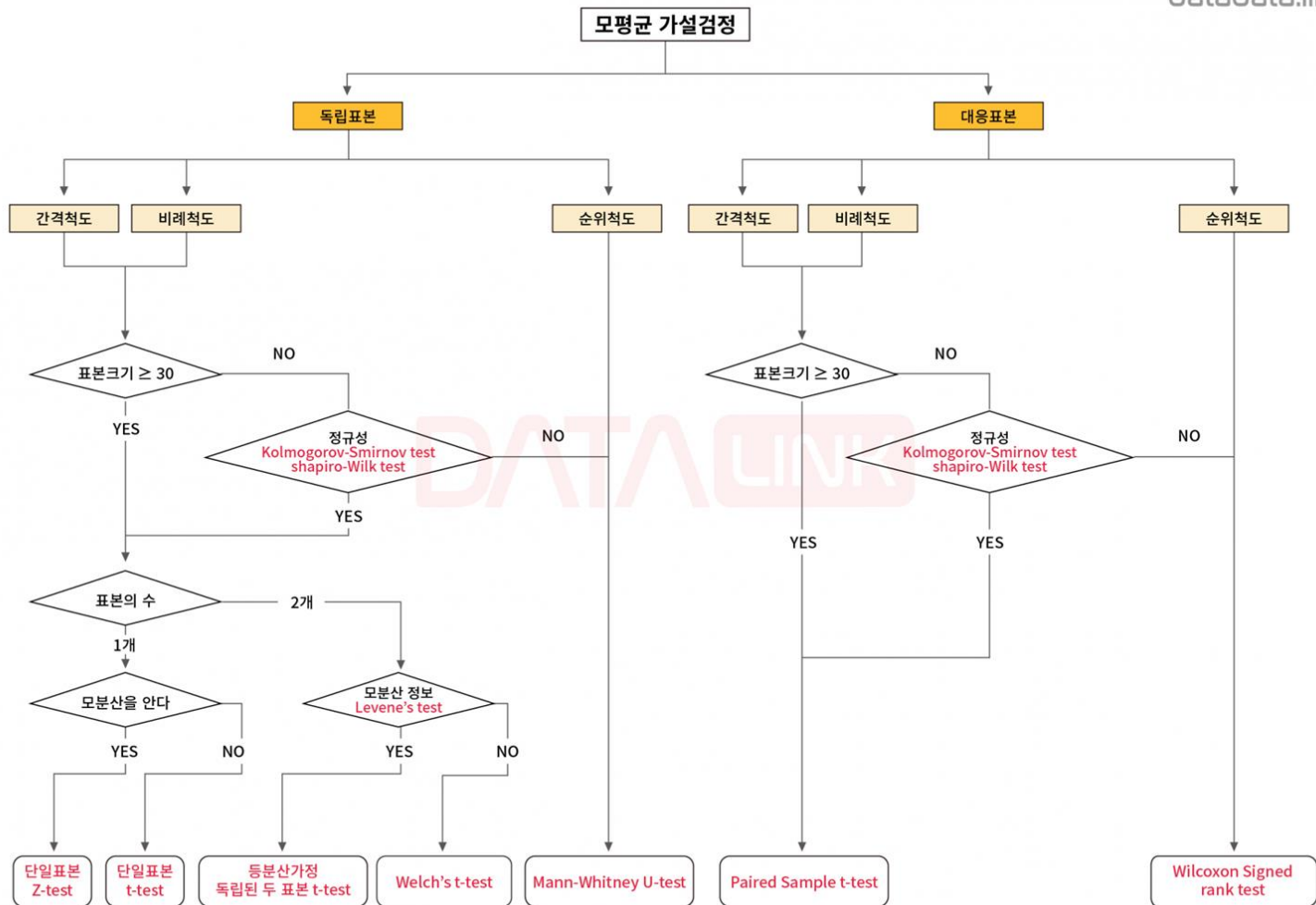
통계로 이상적인 흠의 비율을 찾아낸다.

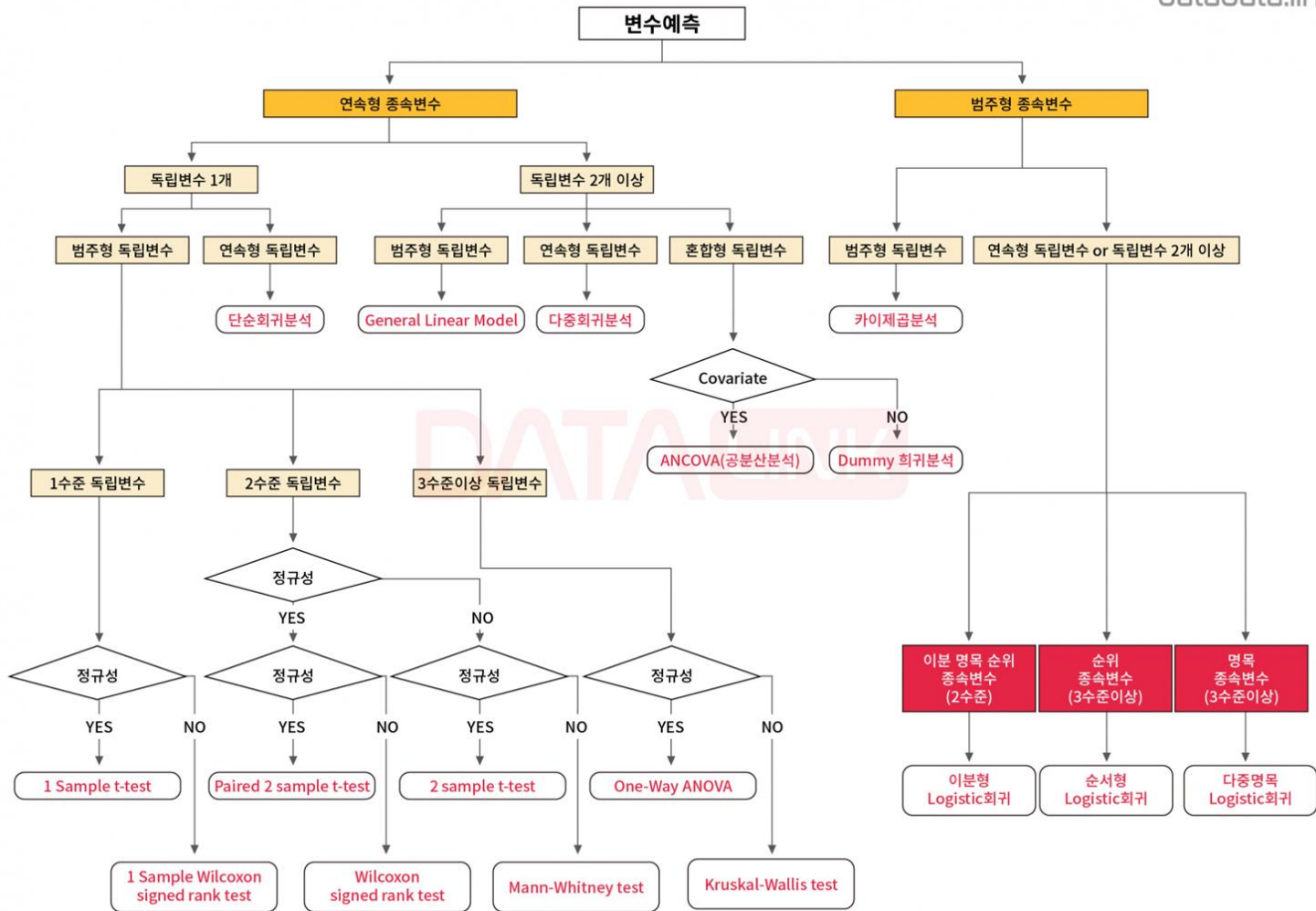


[William Sealy Gosset](#), who developed the "t-statistic" and published it under the [pseudonym](#) of "Student".

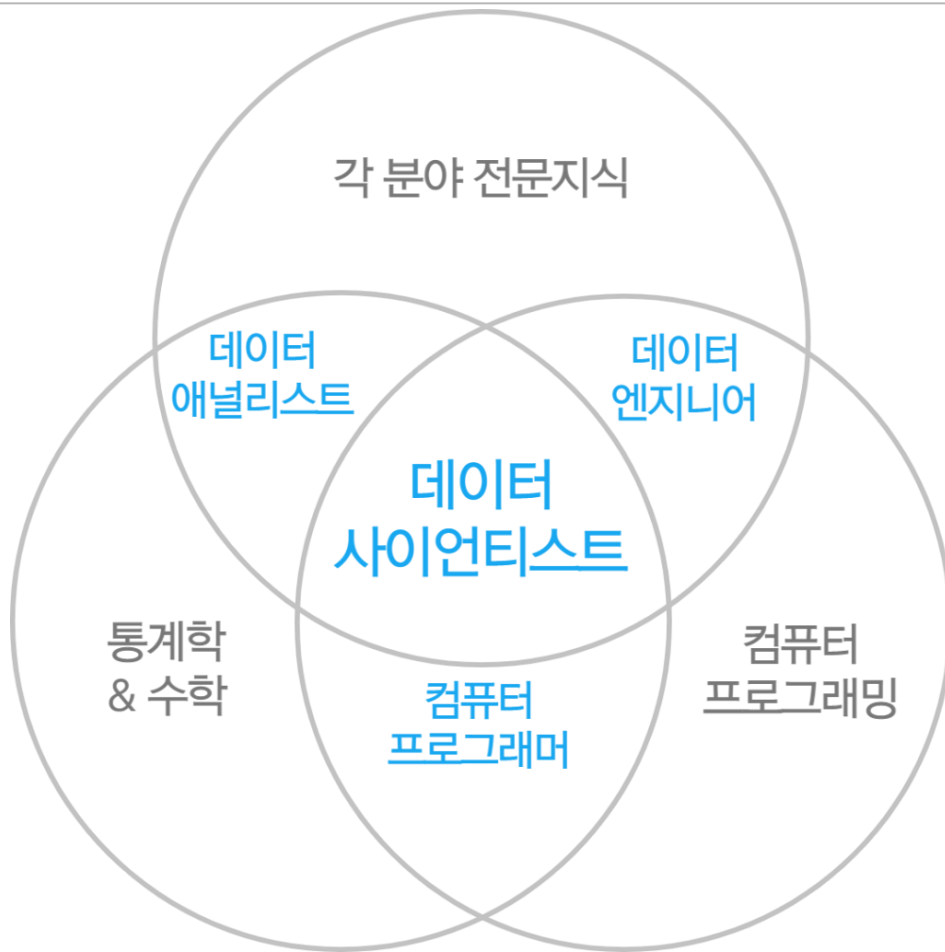
Normal Distribution(정규분포)







통계, 데이터사이언스, AI, Machine Learning



Thank You
