

통계적 사고

23.01.12 / 8기 정건우

CONTENTS

01. Why statistics?

- 숫자로 말하는 방법
- 분석대상(표본)을 고르는 방법
- 무엇보다도

02. t-test

- t-test란?
- t-test에 대한 이해
- 자유도
- t-test의 종류

03. Anova

- One-way ANOVA
- F-value
- Two-way ANOVA

04. Regression

- Regression line
- Regression and t-test

05. Sampling theory

- What is sampling
- Why sampling
- Rejection sampling
- Reservoir sampling

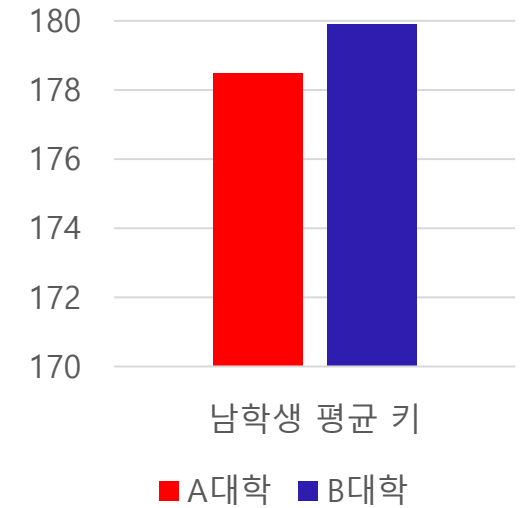
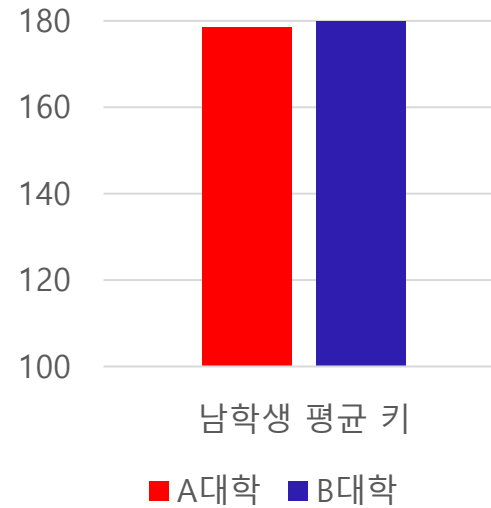
06. Summary

- Summary
- Reference

1. Why statistics?

숫자로 말하는 방법

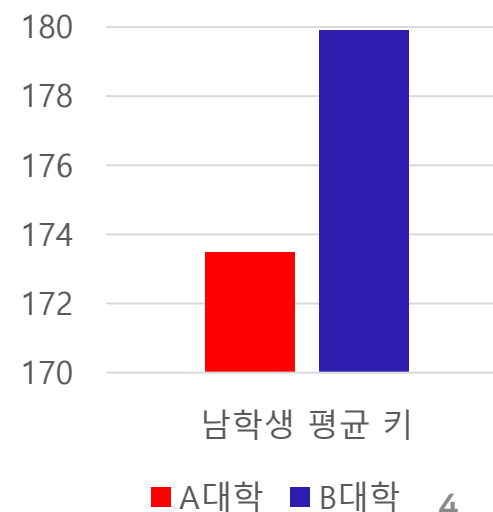
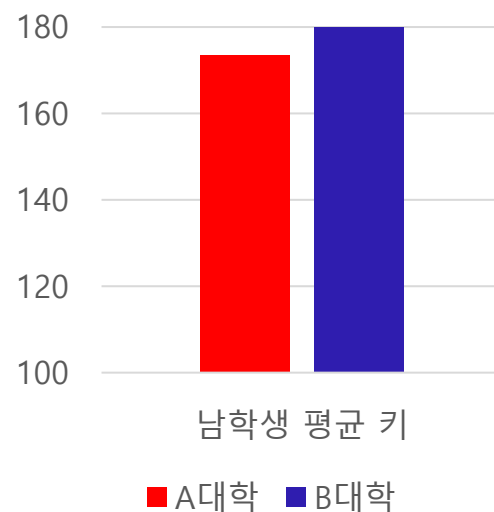
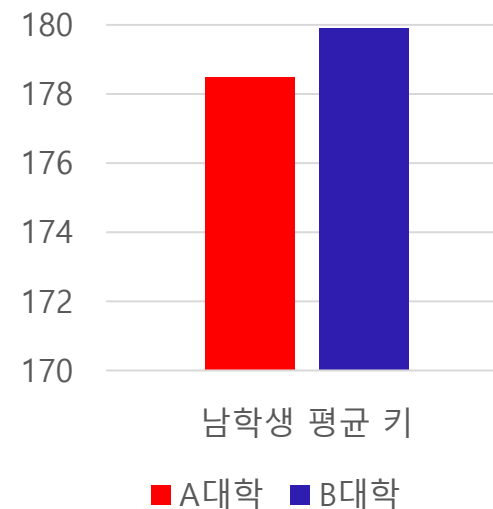
- A대학 남학생 평균 키 178.5cm
B대학 남학생 평균 키 179.9cm
- 두 학교 남학생의 평균 키 차이는 1.4cm
- 두 학교 남학생의 평균 키는 같다? 다르다?



1. Why statistics?

숫자로 말하는 방법

- A대학 남학생 평균 키 178.5cm
B대학 남학생 평균 키 179.9cm
- 두 학교 남학생의 평균 키 차이는 1.4cm
- 두 학교 남학생의 평균 키는 같다? 다르다?

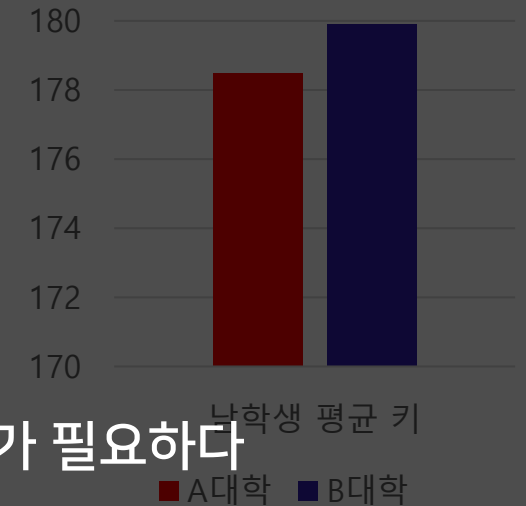
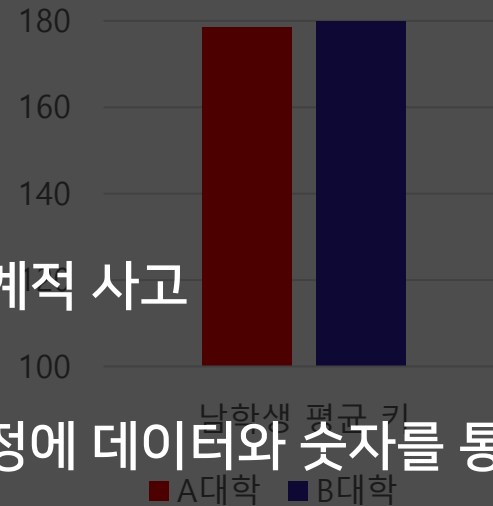


1. Why statistics?

숫자로 말하는 방법

- A대학 남학생 평균 키 178.5cm
B대학 남학생 평균 키 179.9cm
- 두 학교 남학생의 평균 키 차이는 1.4cm
- 이유 1. 프로젝트에서 내리는 의사결정에 데이터와 숫자를 통한 근거가 필요하다
- A대학 남학생 평균 키 173.5cm
B대학 남학생 평균 키 179.9cm
- 두 학교 남학생의 평균 키 차이는 6.4cm
- 두 학교 남학생의 평균 키는 같다? 다르다?

통계적 사고



1. Why statistics?

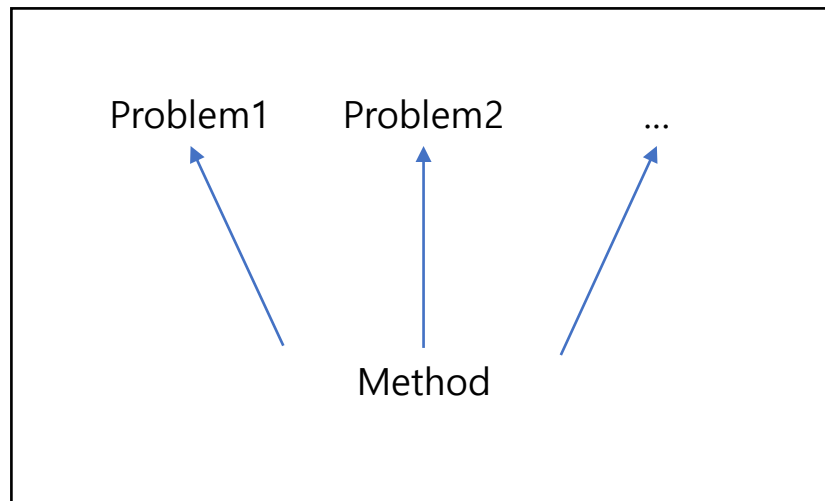
분석 대상(표본)을 고르는 방법

- 모집단을 모두 관찰하기 어려울 때, 표본의 특성을 통해 모집단의 특성을 유추
- 표본을 통해 의사결정을 하더라도, 의사결정의 영향은 모집단 전체에 준다
- 표본이 모집단의 특성을 잘 포함하도록 표본을 뽑을 필요가 있다
- 가진 데이터의 분포가 쉽게 다루기 힘든 분포인 경우 (샘플을 만들기 어려운 경우)
다양한 통계적 기법을 통해, 쉬운 방법으로 샘플을 만들더라도 (ex. Uniform, Normal 분포)
우리가 가진 데이터의 복잡한 분포로부터 샘플을 만든 것'처럼' 간주하고 분석할 수 있다

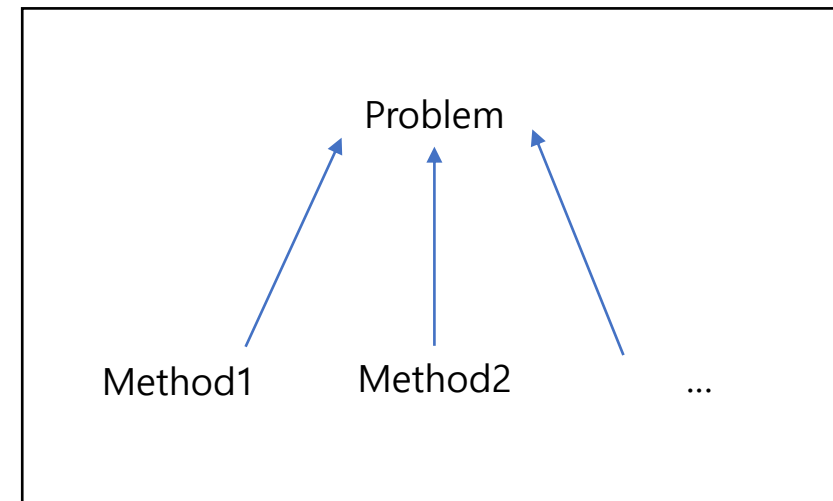
1. Why statistics?

무엇보다도

- “If your only tool is a hammer, every problem looks like a nail”



사회과학



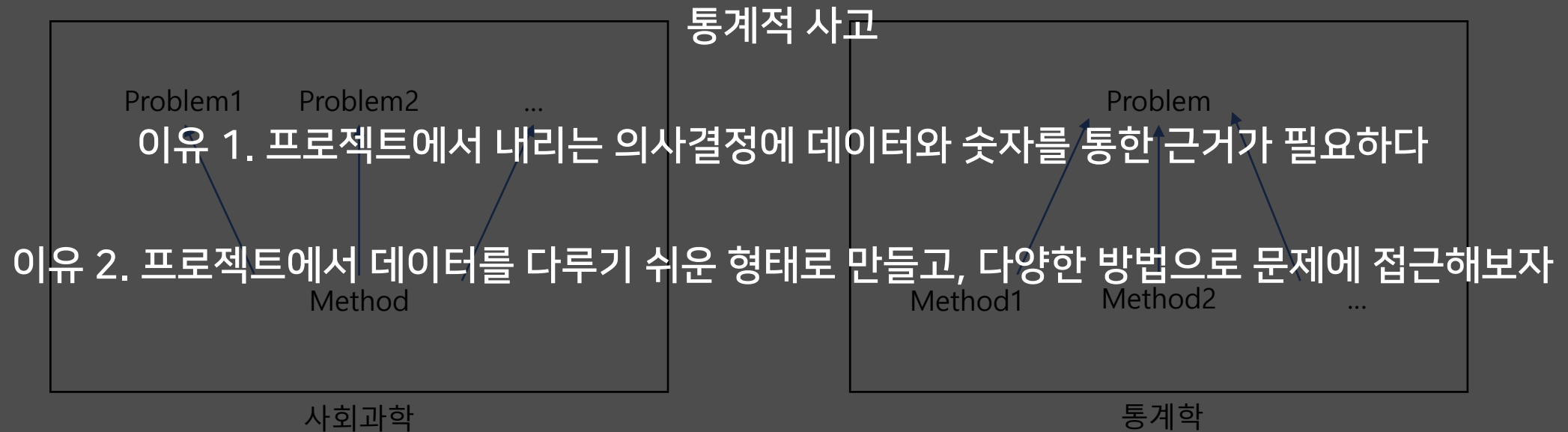
통계학

- 사회과학은 다양한 문제를 해결할 수 있는 method를 탐구하는 학문이라면,
통계학은 하나의 문제를 어떤 method로 해결해야 가장 잘 해결할 수 있을지 탐구하는 학문

1. Why statistics?

무엇보다도

- “If your only tool is a hammer, every problem looks like a nail”



- 사회과학은 다양한 문제를 해결할 수 있는 method를 탐구하는 학문이라면,
통계학은 하나의 문제를 어떤 method로 해결해야 가장 잘 해결할 수 있을지 탐구하는 학문

2. t-test

t-test란?

- 모집단의 표준편차가 알려지지 않았을 때, 정규분포의 모집단에서 모든 샘플의 평균값에 대한 가설검정 방법

왜 t-test인가?

- 1908년 영국의 William Sealy Gosset이 개발한 방법
- 당시 William Sealy Gosset의 가명이 Student여서 마지막 글자 't'를 따왔다...

2. t-test

t-test의 목적

- 두 표본(sample)의 평균이 같다 or 다르다 비교하기 위해 사용
- 예시)
A대학 남학생 평균 키 178.5cm
B대학 남학생 평균 키 179.9cm
- A, B대학 남학생 평균 키 차이 1.4cm가 통계적으로 유의미할까? (우연일까?)
- 1.4cm가 유의미한 차이인지 판단하는 기준(비교대상)이 표준편차

t-test에 대한 이해

- 왜 표준편차가 비교대상일까? 다시 처음부터

$$[1, 2, 3, 4, 5] \text{의 분산} = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{4} = 2.5 \text{ 이고, 표준편차} = \sqrt{2.5} \approx 1.58$$

- 표준편차의 의미는

데이터는 평균값 3을 기준으로 평균적으로 1.58만큼 퍼져있다.

즉, 데이터에 큰 문제가 없는 한, 의미없이 우연히 퍼져있는 정도

- 예시)

A, B대학 남학생 평균 키 차이 1.4cm가 A,B 데이터의 표준편차보다 현저히 크면 -> 1.4cm는 유의미하다

A, B대학 남학생 평균 키 차이 1.4cm가 A,B 데이터의 표준편차보다 현저히 작으면 -> 1.4cm는 의미가 없다.

- t-test는 두 표본의 평균의 차이와, 표준편차의 비율이 얼마나 큰지 작은지를 보는 통계적 과정

2. t-test

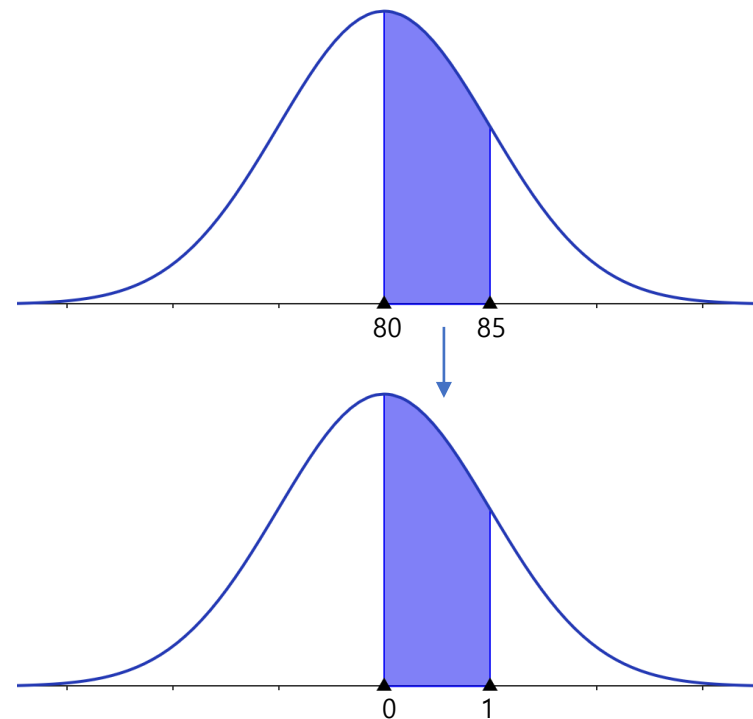
t-test 들어가기 전에 z-test

- t-test를 이해하기 위해서는 정규분포의 z-test에 대한 이해가 필요
- 예시)

학생 1,000명의 점수의 평균이 80이고 표준편차가 5일 때,
80점에서 85점인 학생들의 수를 알기 위해서는 정규분포를 적분해야 함
→ 매번 다르게 생긴 정규분포를 적분하기보다는,
표준정규분포 $N(0,1)$ 로 바꿔서 적분은 표준정규분포에서만 하자

- $z\text{-value} = \frac{X - \mu}{\sigma}$ (standardization)
- $Z_{(90)} = \frac{85 - 80}{5} = 1$
- $0.3413 * 1,000 \approx 341\text{명}$

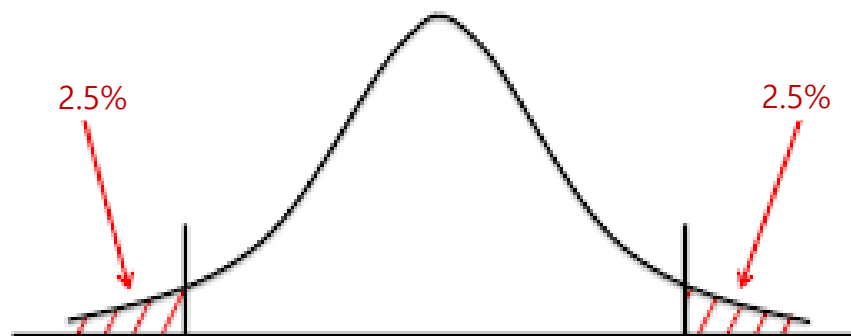
Z	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.1	0.5398	0.5438	0.5478	0.5517	0.5557
0.2	0.5793	0.5833	0.5873	0.5912	0.5951
0.3	0.6179	0.6217	0.6255	0.6293	0.6331
0.4	0.6554	0.6591	0.6628	0.6664	0.6700
0.5	0.6915	0.6950	0.6985	0.7019	0.7054
0.6	0.7257	0.7291	0.7324	0.7357	0.7389
0.7	0.7580	0.7613	0.7643	0.7673	0.7703
0.8	0.7881	0.7910	0.7939	0.7967	0.7995
0.9	0.8159	0.8186	0.8212	0.8238	0.8264
1.0	0.8413	0.8438	0.8461	0.8485	0.8508
1.1	0.8643	0.8665	0.8686	0.8708	0.8729
1.2	0.8849	0.8869	0.8888	0.8907	0.8925



2. t-test

t-test를 위한 가설검정 (양측검정)

- $H_0 : \overline{X}_a = \overline{X}_b \rightarrow H_0 : \overline{X}_a - \overline{X}_b = 0$
 $H_1 : \overline{X}_a \neq \overline{X}_b$



- A, B대학 남학생 평균 키 차이 1.4cm가 **기각역**에 들어가면 우연이라고 보기 어렵다
- A, B대학 남학생 평균 키 차이 1.4cm가 **기각역**에 들어가지 않으면
우연히 발생한 차이이므로 두 표본의 평균값은 통계적으로는 같다

2. t-test

t-value의 의미

- $z\text{-value} = \frac{X - \mu}{\sigma}$
- $t\text{-value} = \frac{\overline{X}_a - \overline{X}_b}{s/\sqrt{n}}$

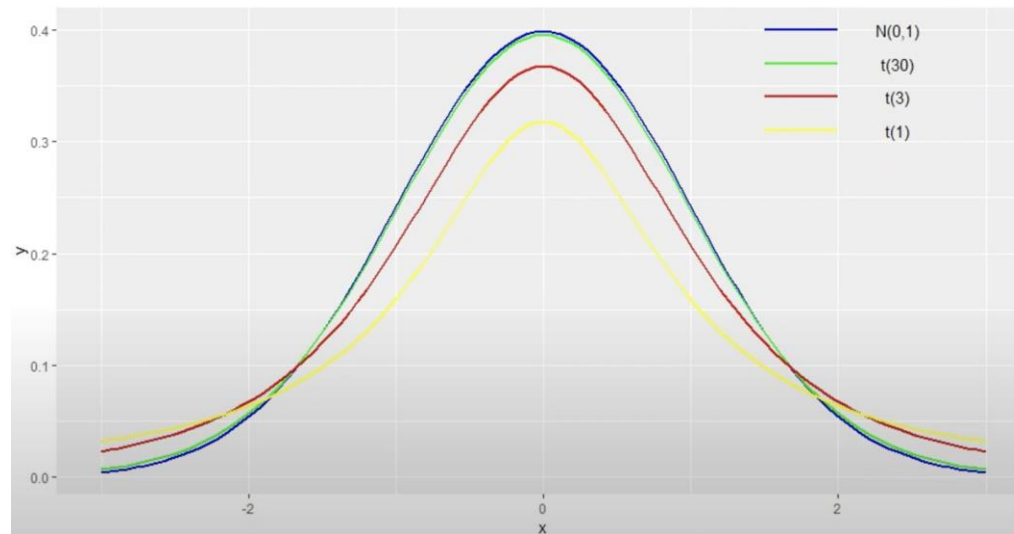
df(degree of freedom): 자유도

- $\overline{X}_a - \overline{X}_b$: 두 집단의 평균의 차이
- s/\sqrt{n} : 데이터가 평균값을 기준으로 평균적으로 퍼진 정도 → 의미 없는 편차
: 두 집단의 평균의 차이가 유의미하게 크다 or 우연이다 판단할 비교대상
(표본의 표준편차는 표본의 개수 n와 관련)

2. t-test

자유도

- 표본의 크기 n 은 자유도 $n-1$ 을 결정 $\rightarrow n$ 이 커질수록 t분포가 표준정규분포에 근사



- $t\text{-value} = \frac{\overline{X_a} - \overline{X_b}}{s/\sqrt{n}}$
- 표본의 크기가 커져 자유도가 커지면, t-분포를 쓰도록 제한되다가 자유롭게 표준정규분포 사용할 수 있다 15

2. t-test

t-test 예제

- 예시)

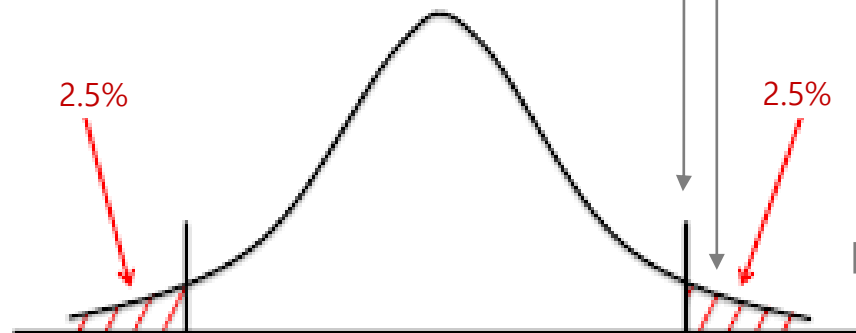
A대학 남학생 평균 키 178.5cm

B대학 남학생 평균 키 179.9cm

- 만약 표준편차(s)가 7.05cm, 표본의 크기(n)가 101명이라면

$$t\text{-value} = \frac{\bar{X}_a - \bar{X}_b}{s/\sqrt{n}} \approx 1.996$$

- Critical value = 1.984



cum. prob	t.50	t.75	t.80	t.85	t.90	t.95	t.975
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960
	0%	50%	60%	70%	80%	90%	95%
Confidence Level							

평균의 차이 1.4cm가 우연히 발생했을 확률이 5%보다 작으니 이 차이는 통계적으로 유의(유의미)하다.

2. t-test

t-test 예제

- 예시)

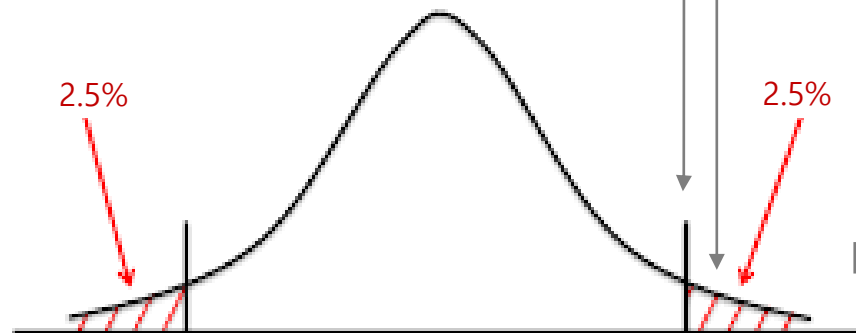
A대학 남학생 평균 키 178.5cm

B대학 남학생 평균 키 179.9cm

- 만약 표준편차(s)가 7.05cm, 표본의 크기(n)가 101명이라면

$$t\text{-value} = \frac{\bar{X}_a - \bar{X}_b}{s/\sqrt{n}} \approx 1.996$$

- Critical value = 1.984



p-value : probability value. 어떤 사건이 우연이 발생할 확률

cum. prob	t.50	t.75	t.80	t.85	t.90	t.95	t.975
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960
	0%	50%	60%	70%	80%	90%	95%
Confidence Level							

평균의 차이 1.4cm가 우연히 발생했을 확률이 5%보다 작으니 이 차이는 통계적으로 유의(유의미)하다.

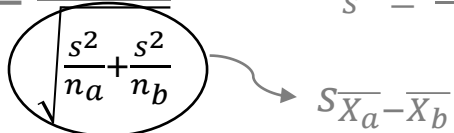
2. t-test

t-test 종류

- Two-sample t-test
 - 서로 다른 두 표본을 비교
 - 예시)
A대학 남학생 평균 키 178.5cm
B대학 남학생 평균 키 179.9cm
통계적으로 같다?
- One-sample t-test
 - 하나의 표본을 검정
 - 예시)
A대학 남학생 평균 키 178.5
통계적으로 180cm 이다?
- Paired-sample t-test
 - 하나의 표본의 두 시점 비교
 - 예시)
A대학 2019년 점수 76.4점
A대학 2020년 점수 84.1점
통계적으로 동일하다?

- t-value는 정확히 말하면 $\frac{\bar{X}_a - \bar{X}_b}{s/\sqrt{n}}$ 가 아니라

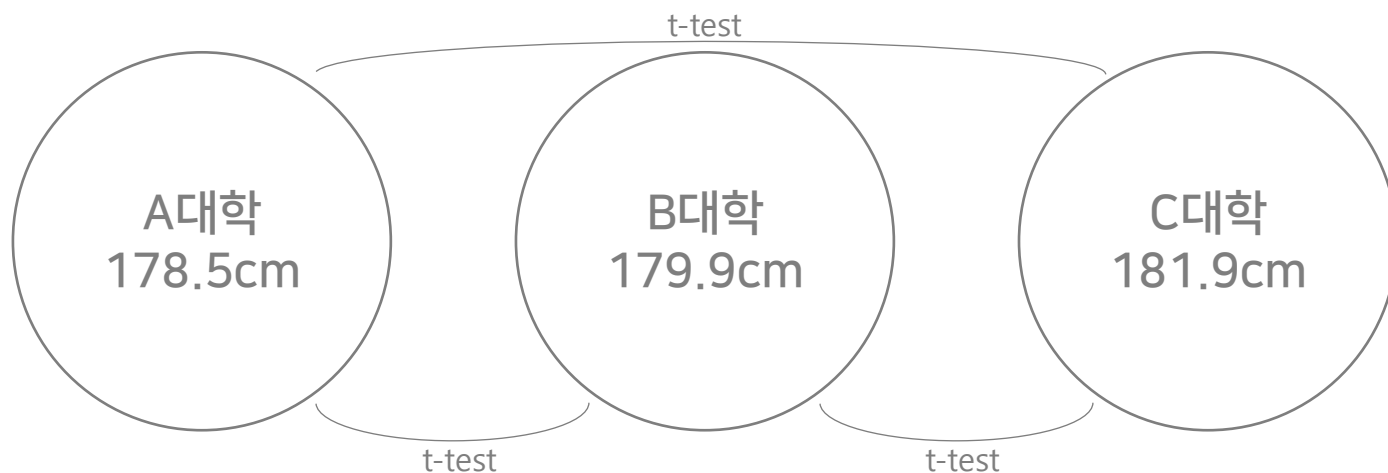
$$t\text{-value} = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{s^2}{n_a} + \frac{s^2}{n_b}}} \quad s^2 = \frac{\sum(x_a - \bar{X}_a)^2 + \sum(x_b - \bar{X}_b)^2}{n_a + n_b - 2}, \quad df = n_a + n_b - 2$$



3. ANOVA

t-test란?

- 모집단의 표준편차가 알려지지 않았을 때, 정규분포의 모집단에서 모든 샘플의 평균값에 대한 가설검정 방법
- 표본이 2개보다 많으면?



		진실	
		H_0 참	H_0 거짓
연구 결과	H_0 참	문제없음	2종오류(β)
	H_0 거짓	1종오류(α)	문제없음

- Type I Error : $1 - (1 - a)^c \approx a * c$ $a = 0.05, c = 3$
- Multiple t-test를 하면 1종 오류의 가능성이 높아진다

3. ANOVA

ANOVA란?

- Analysis of Variance (분산분석)

One-way ANOVA란?

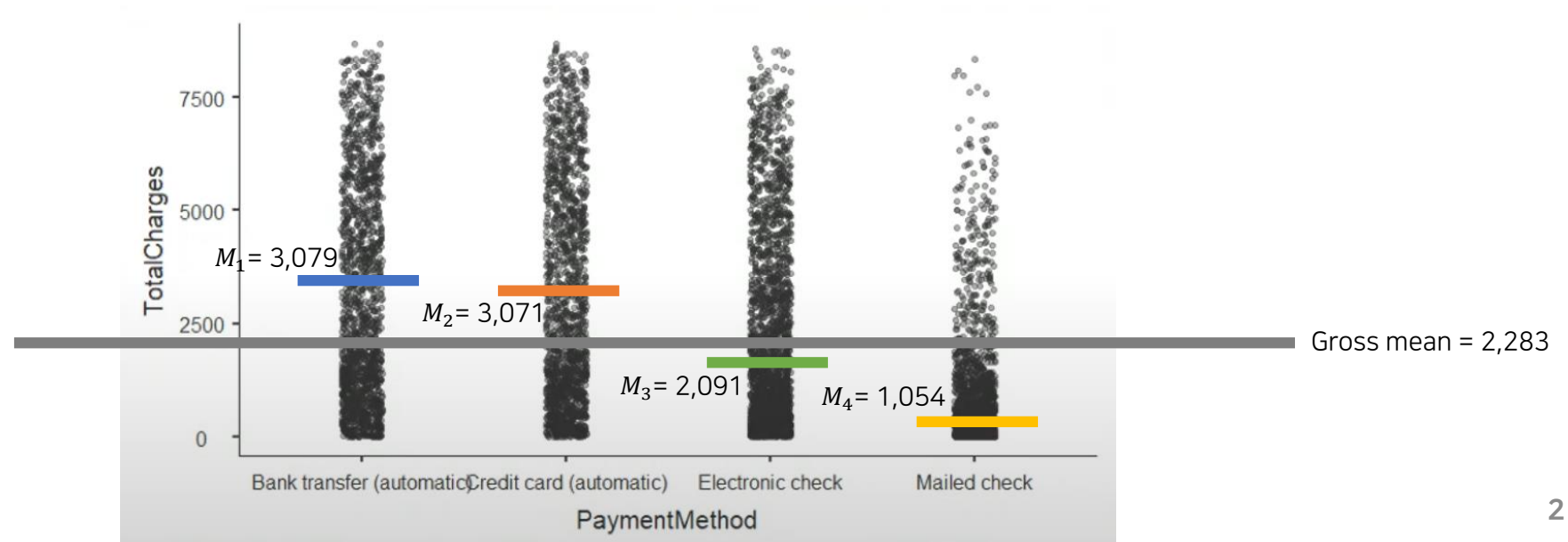
- 하나의 변수(One way)에 대해 세 개 이상의 표본의 평균이 같다 or 다르다 비교하기 위해 사용
- 왜 평균분석이 아니라 분산분석일까?
 - ANOVA에서 사용하는 F-value는 2개의 분산의 비율이기 때문

3. ANOVA

F-value란?

- F-value는 2개의 분산의 비율이므로, 2개의 평균이 필요
- 예시)

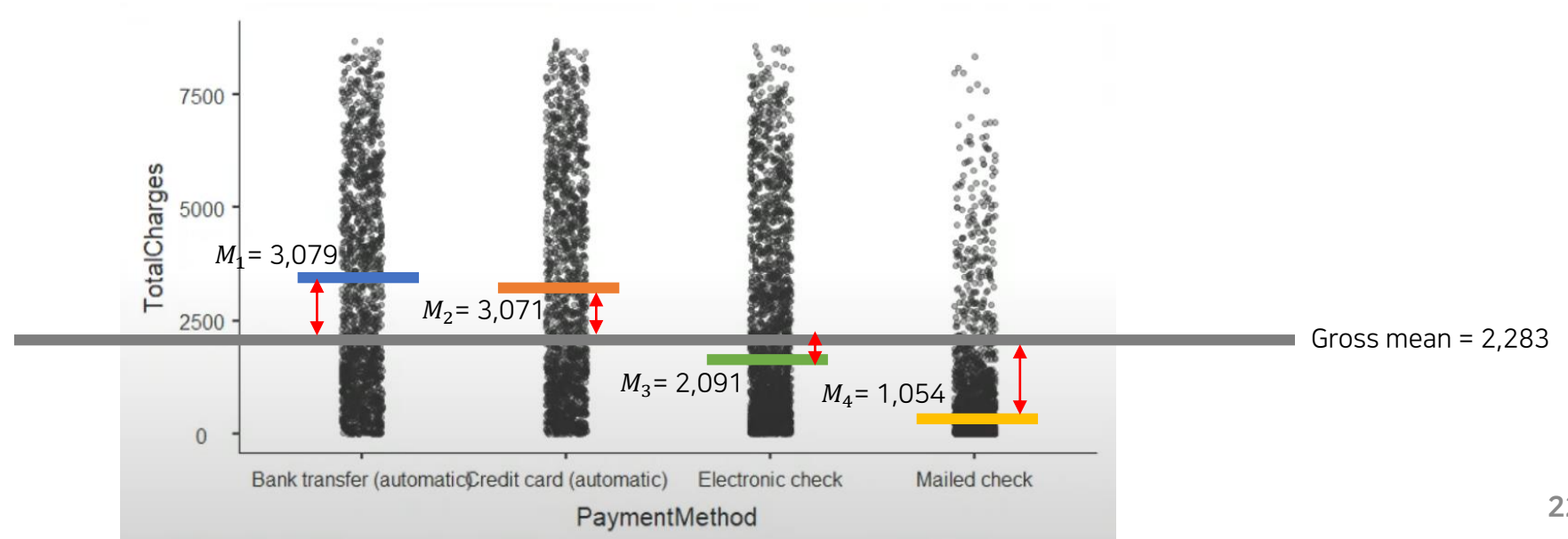
독립변수 : PaymentMethod / 종속변수 : TotalCharges



3. ANOVA

F-value란?

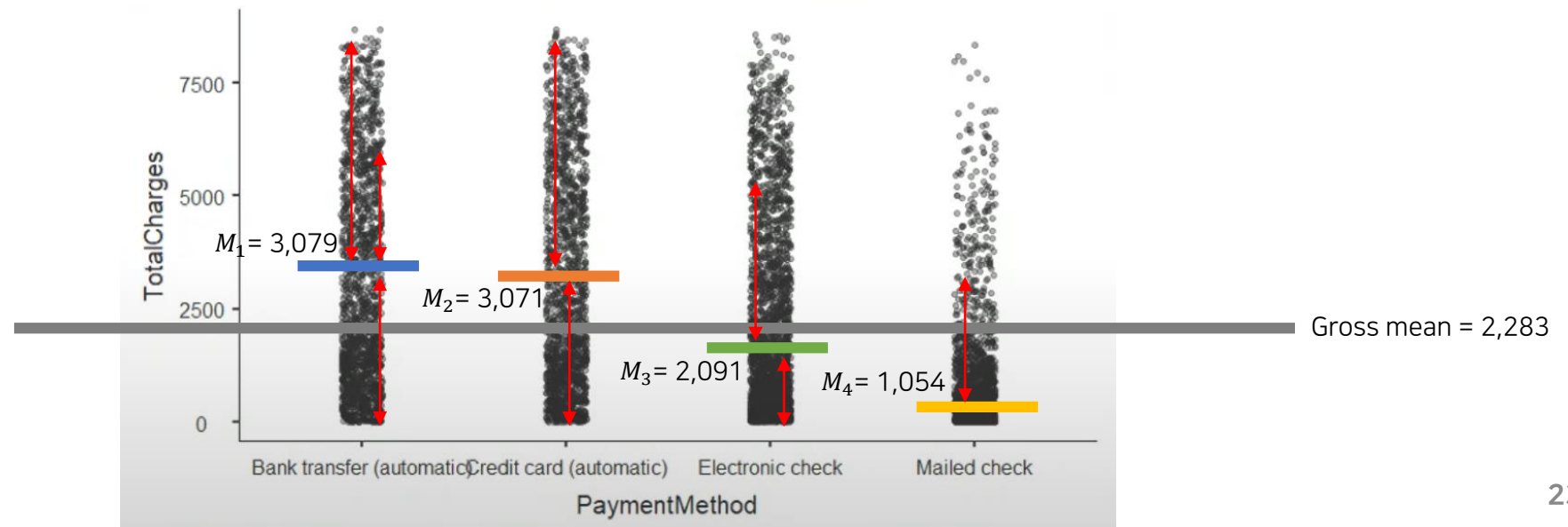
- 첫 번째 분산 : 전체평균(GM)으로부터 각 그룹의 평균 사이의 분산 → **Between Variance**
- Between Variance가 크다는 것은, 적어도 그룹 1개는 다른 그룹과 평균이 다를 수 있음을 의미
- Between Variance가 얼마나 커야 할까? 우연히 큰 값을 가질 확률은 얼마나 될까?
→ 비교 대상 Variance가 필요



3. ANOVA

F-value란?

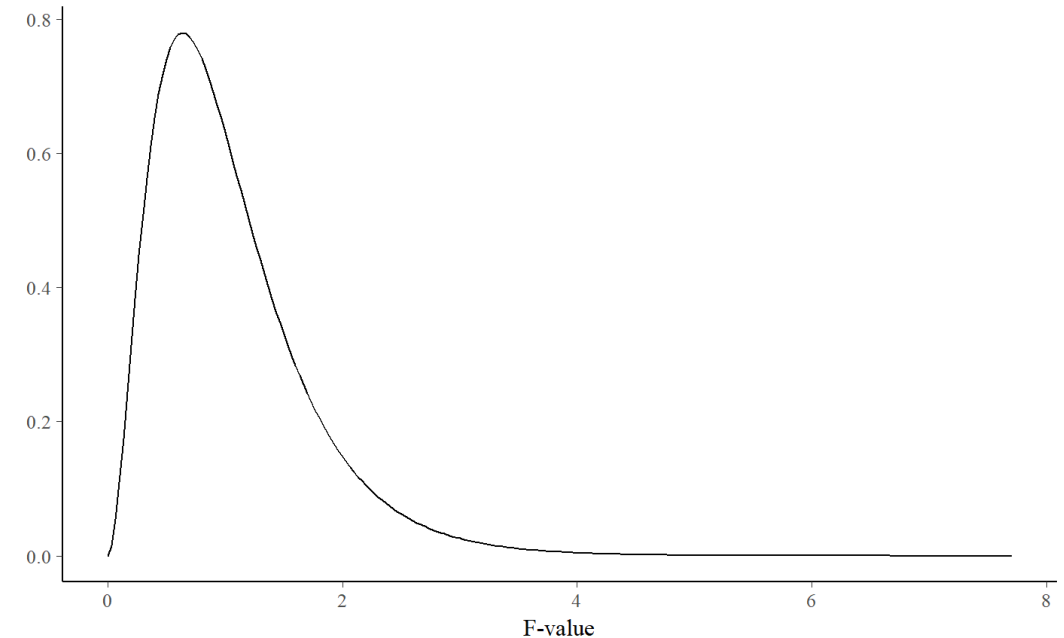
- 두 번째 분산 : 각 그룹 내 분산 → Within Variance
- t-value에서 분모(표준편차)와 같은 의미 → 의미 없는 편차
- Between Variance가 Within Variance보다 충분히 커야
→ Between Variance가 크다고 할 수 있다 → 적어도 한 그룹의 평균이 전체의 평균과 다르다고 할 수 있다



3. ANOVA

F-value란?

- $F\text{-value} = \frac{MS_{Between}}{MS_{Within}}$
- $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ (k:그룹 개수)
 $H_1 : \mu_1 \neq \mu_2$ for some $i, j \rightarrow$ 적어도 한 그룹의 평균은 다르다
- F-value가 충분히 크더라도 (p-value < 0.05),
어떤 그룹이 전체 평균과 얼마나 다른지 알 수 없음



3. ANOVA

F-value 예제

- 예시)

감기약의 효과를 측정하기 위해 감기약A, 감기약B, 플라시보를 세 그룹에게 각각 복용 (총 10명)

감기가 낫는 데까지 걸린 날짜 측정

Id	days	group	group mean
1	5.3	1	6.0
2	6.0	1	6.0
3	6.7	1	6.0
4	5.5	2	5.95
5	6.2	2	5.95
6	6.4	2	5.95
7	5.7	2	5.95
8	7.5	3	7.533
9	7.2	3	7.533
10	7.9	3	7.533

6.44

between
0.194 = $(6.00 - 6.44)^2$
0.194 = $(6.00 - 6.44)^2$
0.194 = $(6.00 - 6.44)^2$
0.240 = $(5.95 - 6.44)^2$
0.240 = $(5.95 - 6.44)^2$
0.240 = $(5.95 - 6.44)^2$
0.240 = $(5.95 - 6.44)^2$
1.195 = $(7.53 - 6.44)^2$
1.195 = $(7.53 - 6.44)^2$
1.195 = $(7.53 - 6.44)^2$

df=3-1=2

5.127

within
0.490 = $(5.30 - 6.00)^2$
0.000 = $(6.00 - 6.00)^2$
0.490 = $(6.70 - 6.00)^2$
0.203 = $(5.50 - 5.95)^2$
0.063 = $(6.20 - 5.95)^2$
0.203 = $(6.40 - 5.95)^2$
0.063 = $(5.70 - 5.95)^2$
0.001 = $(7.50 - 7.53)^2$
0.111 = $(7.20 - 7.53)^2$
0.134 = $(7.90 - 7.53)^2$

df=10-3=7

1.757

F-value

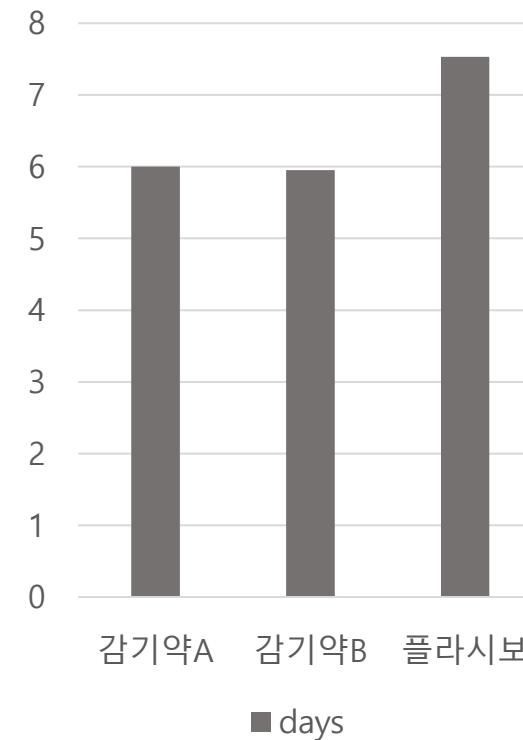
$$\begin{aligned} &= \frac{MS_{\text{Between}}}{MS_{\text{Within}}} \\ &= (5.127/2)/(1.757/7) \\ &= 10.216 \end{aligned}$$

3. ANOVA

F-value 예제

- $$F\text{-value} = \frac{MS_{Between}}{MS_{Within}} = (5.127/2)/(1.757/7) = 10.216$$


F-table of Critical Values of $\alpha = 0.05$ for F(df1, df2)														
	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74





- F-value가 critical value 4.74보다 크므로,
p-value는 0.05보다 작고, 세 감기약의 차이는 유의하다
- 어떤 그룹(감기약)이 얼마나 다를까? → 사후검정 (Bonferroni / Sheffe / Turkey ...)

3. ANOVA

ANOVA table



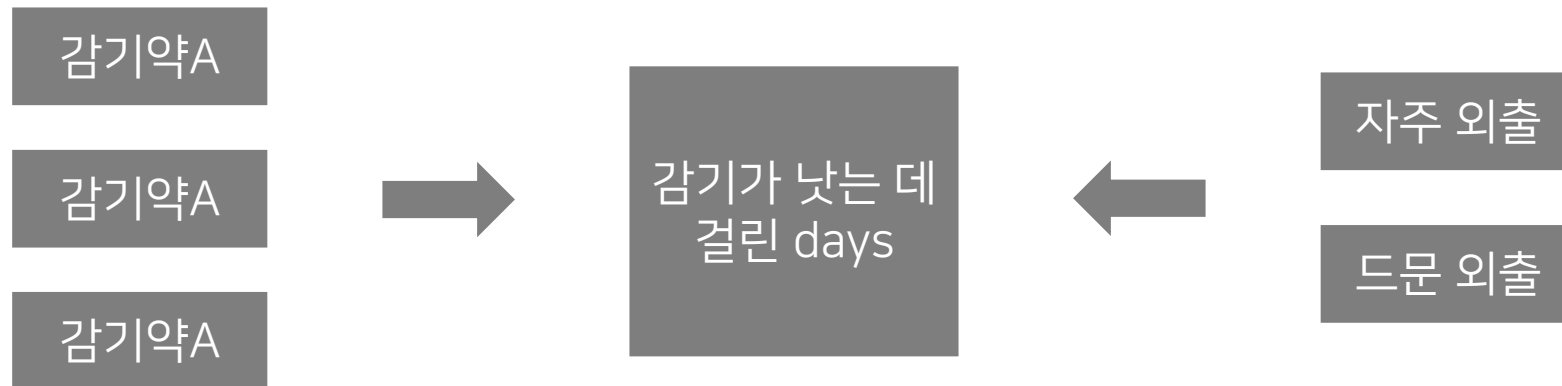
```
model = ols('rand ~ C(level)', df).fit()
anova_lm(model)
```

	df	sum_sq	mean_sq	F	PR(>F)	
C(level)	2.0	829.233975	414.616987	195.226173	7.478367e-55	
Residual	297.0	630.761968	2.123778	NaN	NaN	

3. ANOVA

Two-way ANOVA

- t-test → ANOVA 는 하나의 독립변수에 대해 비교그룹이 2개 → 3개일 때
- One-way ANOVA → Two-way ANOVA는 독립변수가 1 → 2개일 때



- Interaction effect : 독립변수1이 종속변수에 주는 영향이, 독립변수2에 따라 변하는 경우
- F-value가 3개 필요 (독립변수1, 독립변수2, Interaction effect)

3. ANOVA

Two-way ANOVA

- $F\text{-value} = \frac{MS_{Between}}{MS_{Within}}$
- MS_{Within} 은 $MS_{Between}$ 가 크다 or 작다 판단하는 비교 대상이므로 3개의 F-value에서 동일해야 함
- $MS_{Between}$ 은 $MS_{Between_1}$, $MS_{Between_2}$, $MS_{Between_{interaction}}$ 3개를 사용

Days	감기약A		감기약B		플라시보		평균
드문 외출	4	5	7	8	12	12	8.33
	5		9		13		
	6		8		10		
	5		8		13		
잦은 외출	6	5	10	11	13	13	9.66
	6		12		15		
	4		11		12		
	4		11		12		
평균	5		9.5		12.5		9

3. ANOVA

Two-way ANOVA

- $MS_{Between_1} = 4 * 2 * \frac{\{(5-9)^2 + (9.5-9)^2 + (12.5-9)^2\}}{(3-1)}$
- $MS_{Between_2} = 4 * 3 * \frac{\{(8.33-9)^2 + (9.66-9)^2\}}{(2-1)}$
- $MS_{within} = \frac{\{(4-5)^2 + (5-5)^2 + ... + (12-13)^2\}}{(4-1) * 2 * 3}$

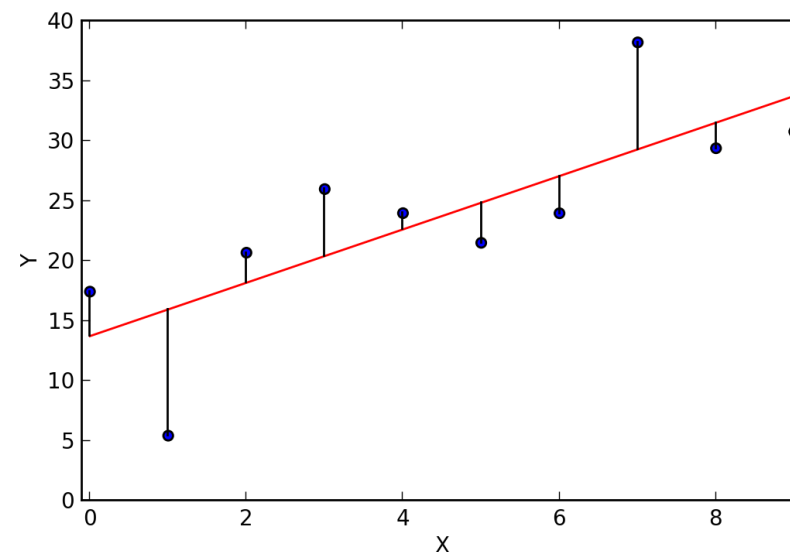
Days	감기약A		감기약B		플라시보		평균
드문 외출	4	5	7	8	12	12	8.33
	5		9		13		
	6		8		10		
	5		8		13		
잦은 외출	6	5	10	11	13	13	9.66
	6		12		15		
	4		11		12		
	4		11		12		
평균	5		9.5		12.5		9

- $MS_{Between_interaction} = 4 * \frac{\{(5-8.33-5+9)^2 + (8-8.33-9.5+9)^2 + ... + (13-9.66-12.5+9)^2\}}{(3-1)*(2-1)}$
- 독립변수1, 독립변수2, interaction effect 모두 종속변수에 주는 영향이 유의하다면 (F-value > critical value)
→ 세 요인 모두에 대한 사후검정이 필요 (Bonferroni / Sheffe / Turkey ...)

4. Regression

What is Regression

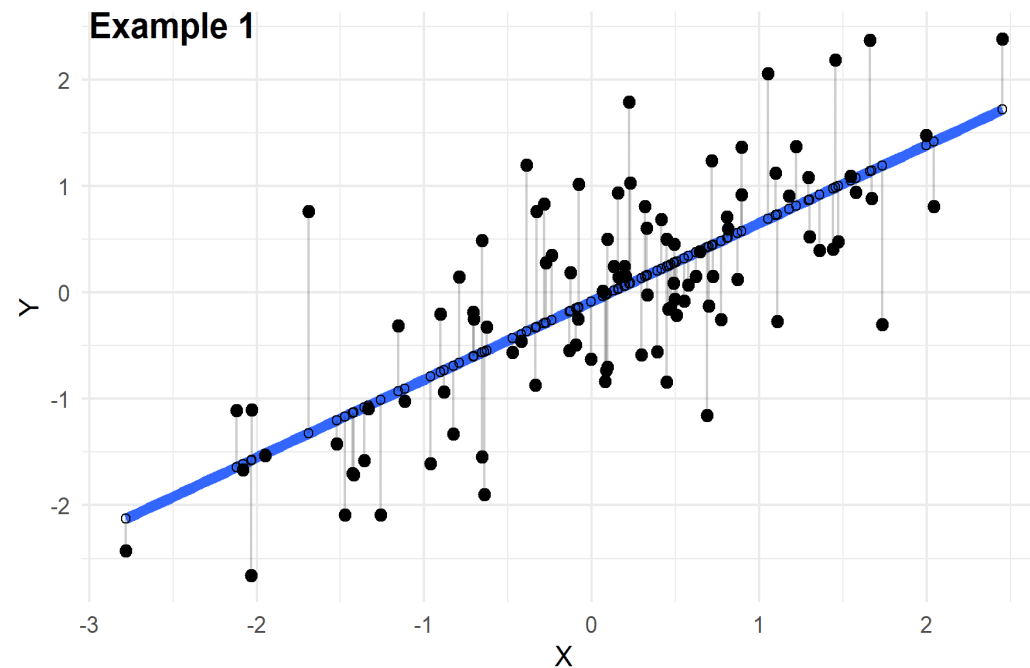
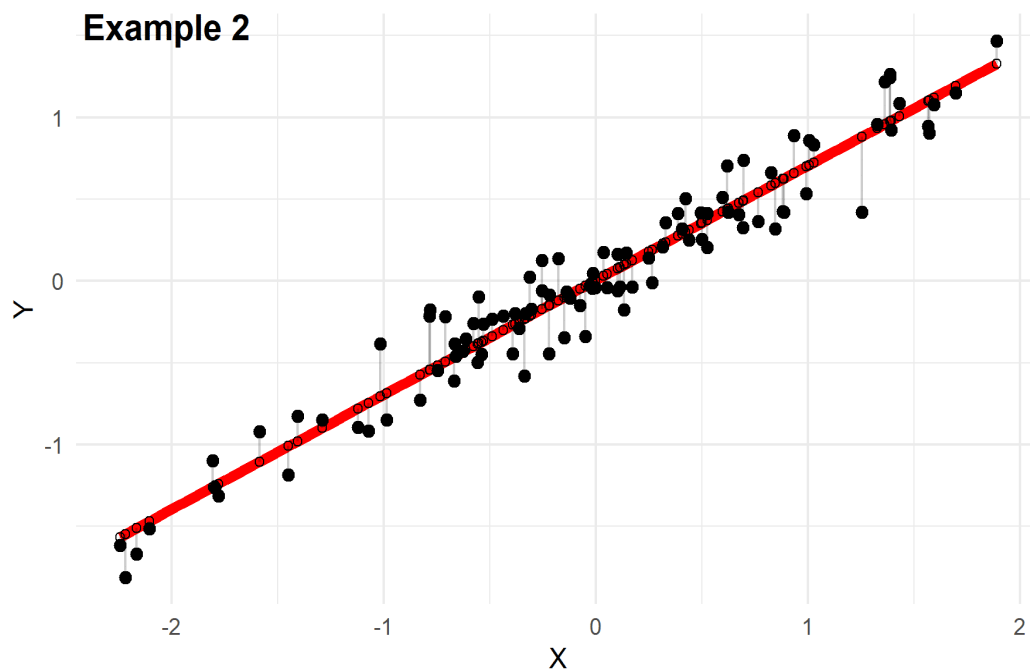
- 독립변수로 종속변수를 예측하는 통계적 방법론
- 추세선 : 데이터를 설명하는 하나의 선 $\hat{y} = a + b\bar{x}$
좋은 추세선 → 오차의 제곱이 최소가 되는 추세선 (OLS)
- 왜 회귀분석일까
→ 추세선이 데이터의 평균을 지나기 때문
→ 평균값으로 회귀한다



4. Regression

Regression line

- 같은 추세선이라 하더라도 유의미한 추세선일 수 있고, 우연한 추세선일 수 있다
 - 비교대상이 필요
 - 표준오차 : 작으면 표본이 모집단의 특성에 가깝고(유의미), 크면 표본이 모집단의 특성과 멀다(우연)



4. Regression

Regression and t-test

- 추세선의 표준오차가 작으면 추세선이 유의미하고, 표준오차가 크면 추세선이 우연에 가깝다
→ t-test로 판단 !

OLS Regression Results

Dep. Variable:	y	R-squared:	0.669
Model:	OLS	Adj. R-squared:	0.667
Method:	Least Squares	F-statistic:	299.2
Date:	Mon, 01 Mar 2021	Prob (F-statistic):	2.33e-37
Time:	16:19:34	Log-Likelihood:	-88.686
No. Observations:	150	AIC:	181.4
Df Residuals:	148	BIC:	187.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	-3.2002	0.257	-12.458	0.000	-3.708	-2.693
x1	0.7529	0.044	17.296	0.000	0.667	0.839

Omnibus:	3.538	Durbin-Watson:	1.279
Prob(Omnibus):	0.171	Jarque-Bera (JB):	3.589
Skew:	0.357	Prob(JB):	0.166
Kurtosis:	2.744	Cond. No.	43.4

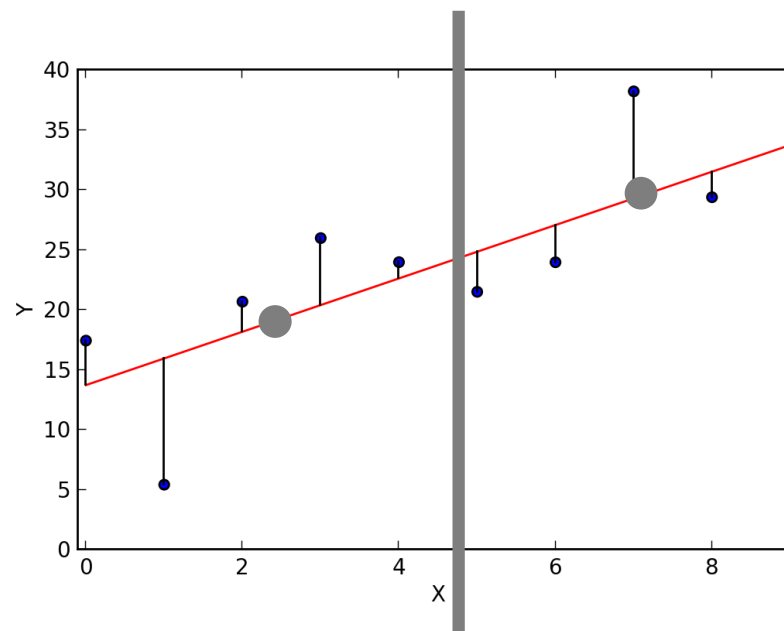
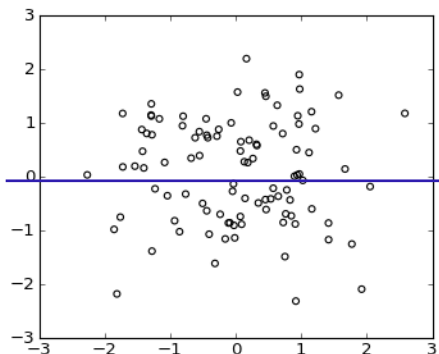
4. Regression

Regression and t-test

- 추세선의 표준오차가 작으면 추세선이 유의미하고, 표준오차가 크면 추세선이 우연에 가깝다
→ t-test로 판단!

- $$t\text{-value} = \frac{\text{평균 집단}_1 - \text{평균 집단}_2}{\text{표준오차}}$$

- 두 집단의 평균이 같다 or 다르다 판단하는 것은
→ 추세선의 기울기가 0인지 판단하는 것과 같다

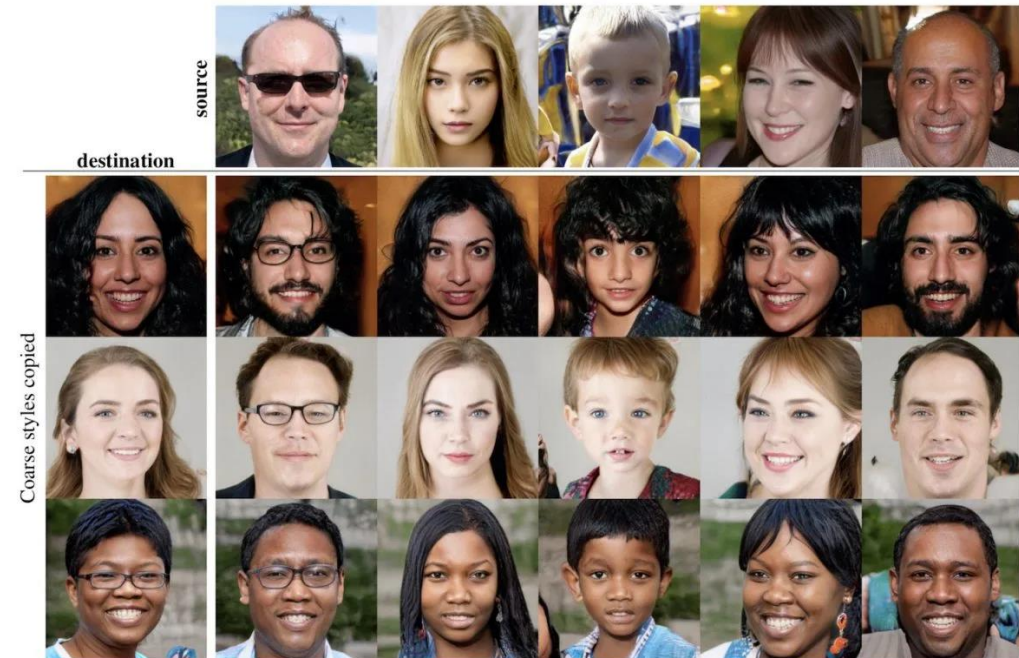
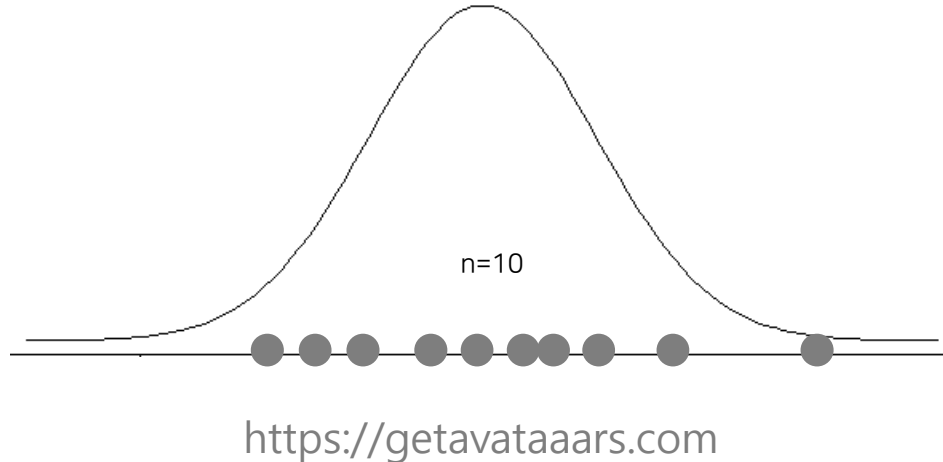


- 회귀분석 전에 산포도(scatter plot)을 그려보아야 한다
- 직선 형태의 데이터 분포가 나타나지 않으면 다른 방법으로

5. Sampling Theory

What is sampling?

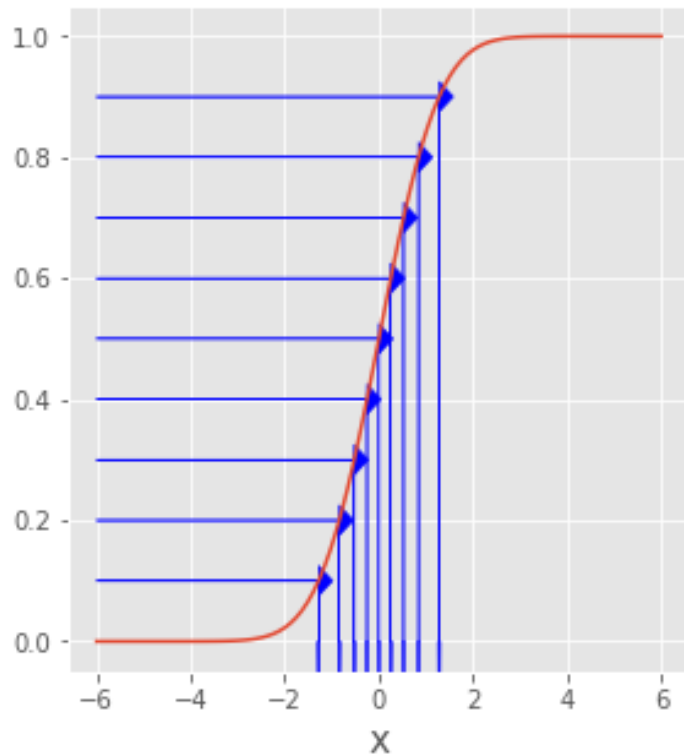
- 일반적인 Sampling : 모집단에서 표본을 추출하는 것
- 통계적인 Sampling : 주어진 확률 분포의 확률 밀도에 맞게 관찰값을 생성하는 것



5. Sampling Theory

What is sampling?

- 수학적 Sampling : CDF(cumulative density function)의 역함수 연산을 수행하는 것



- CDF를 구하기 위해서는 각 샘플이 뽑힐 확률(PDF)을 적분해야 하는데 CDF를 구해도 역함수를 찾는 것이 쉽지 않다
- PDF를 알아도 샘플을 쉽게 추출할 수 있는 것이 아니다 !
- 더 쉽게 Sampling 할 수 있는 방법에 대한 연구
- 세션에서는 수식을 다루지 않음
수식적 이해는 통계 전공 수업에서...

5. Sampling Theory

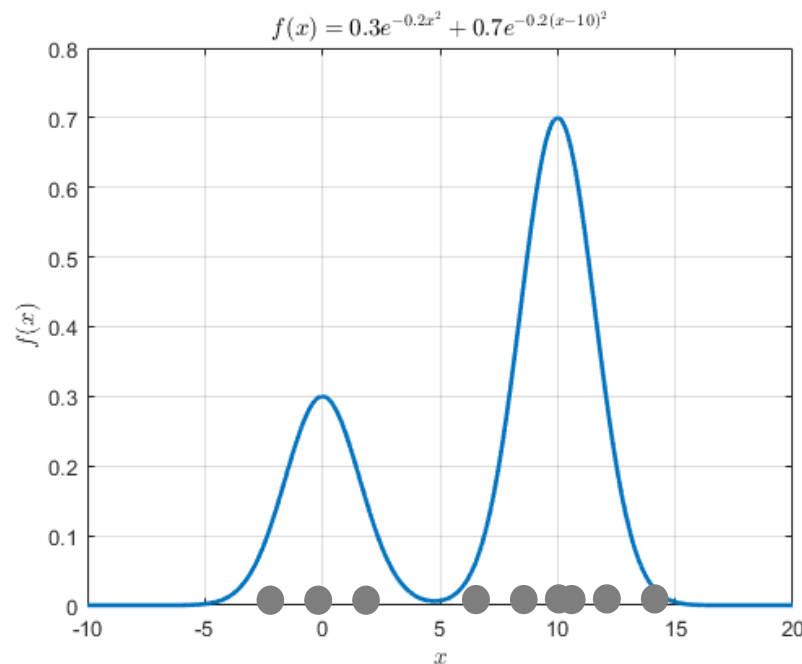
Why sampling?

- 샘플의 목적은 모집단의 특성을 관찰하는 것
- 모집단을 관찰하기에는 시간과 비용이 많이 든다
- 좋은 샘플이란, 모집단의 특성을 잘 반영하고 있는 샘플
- 모집단의 특성을 보다 정확히 알기 위해서는 샘플링을 반복해야 한다
ex. 10,000명 중 200명을 뽑아서 평균을 내는 작업을 100번 해서 구한 100개의 평균
→ 100개의 평균을 다시 평균 내면 → 10,000명의 평균에 수렴
- But 무작위로 샘플을 뽑는 작업만 반복하는 것 역시 시간과 비용이 많이 든다
- 샘플을 한 번 뽑을 때 모집단의 특성이 잘 반영되도록, or 우리가 다루기 쉬운 분포를 따르도록 뽑을 수 있을까?

5. Sampling Theory

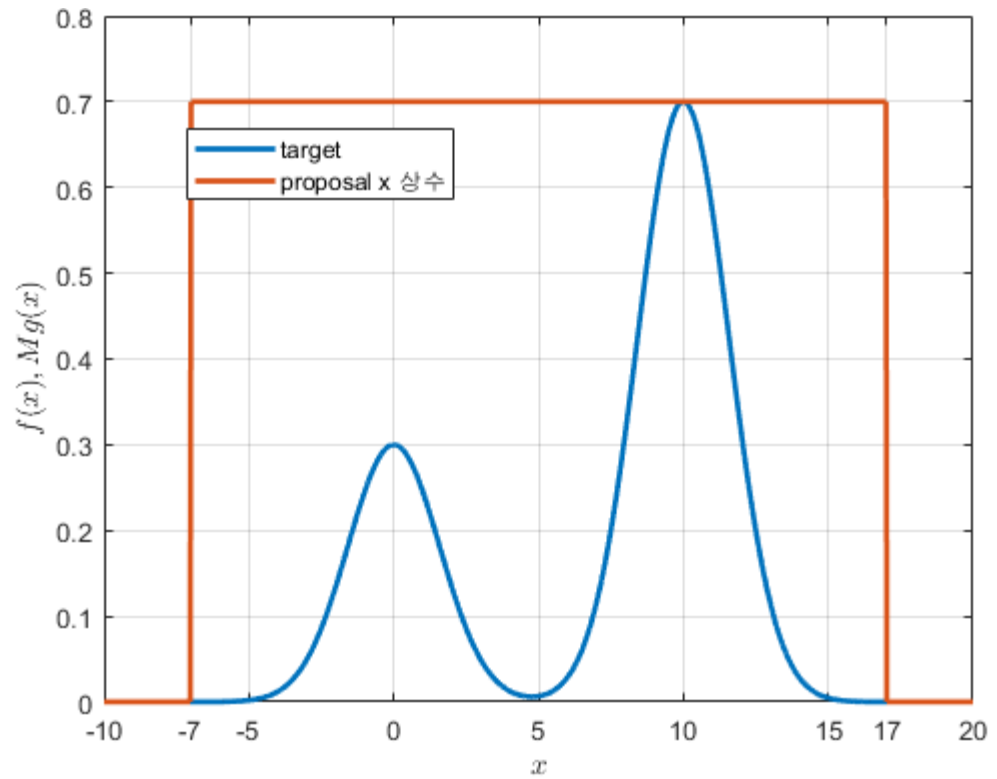
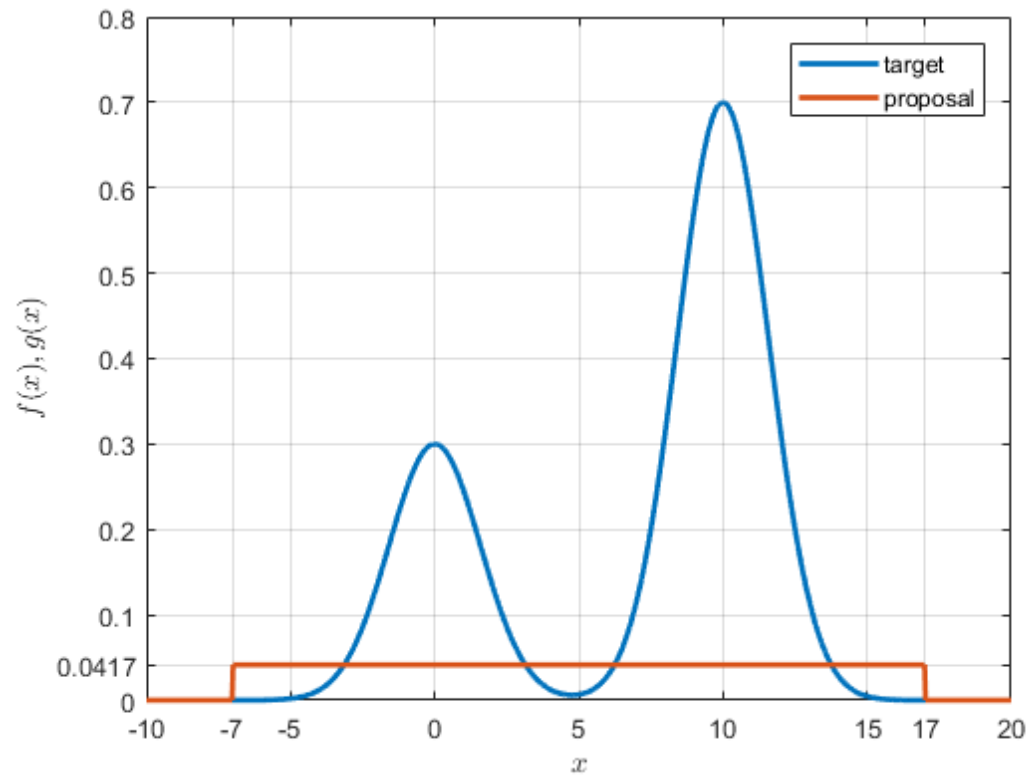
Rejection sampling

- 언제 사용 ? : 샘플을 추출하고자 하는 모집단의 분포에서 샘플을 추출하기 어려울 때
ex. PDF를 적분하여 CDF를 구하기 어렵거나, CDF의 역함수를 구하기 어려울 때
- 예시) $f(x) = 0.3e^{-0.2x^2} + 0.7e^{-0.2(x-10)^2}$ 이 target PDF



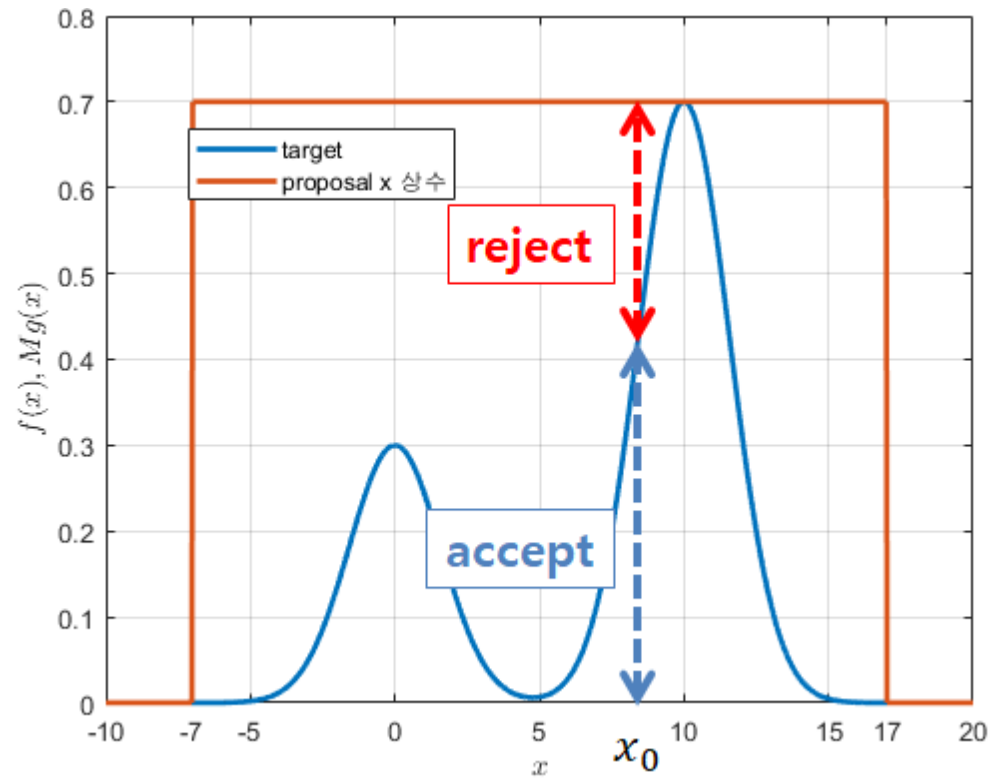
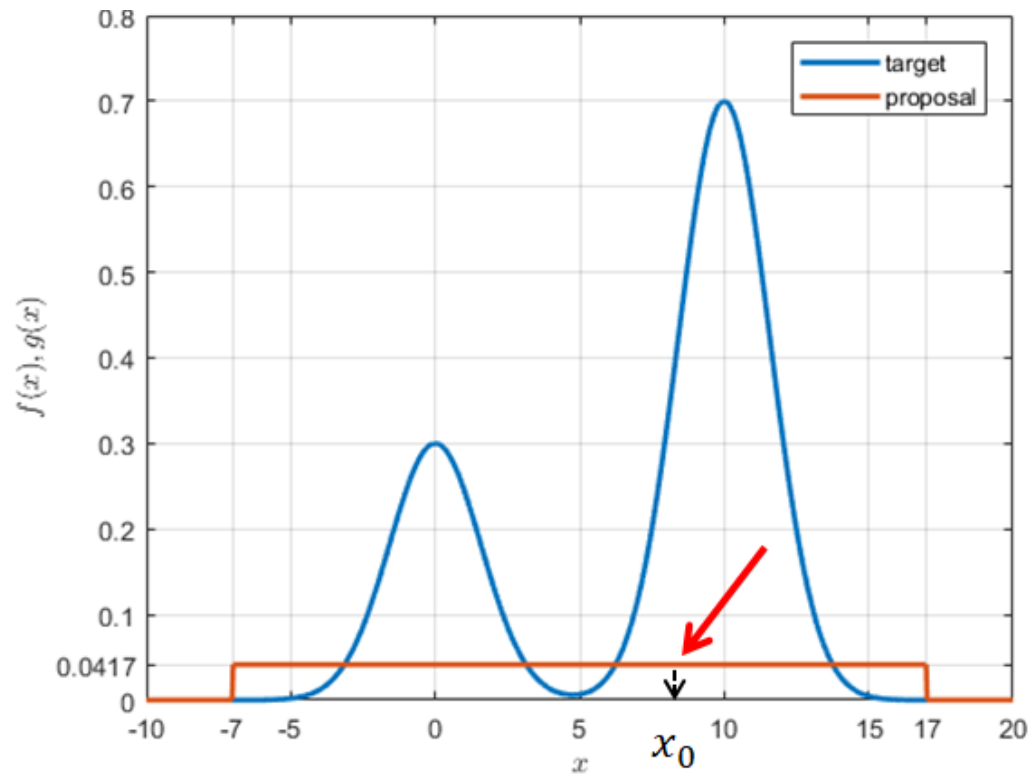
5. Sampling Theory

Rejection sampling



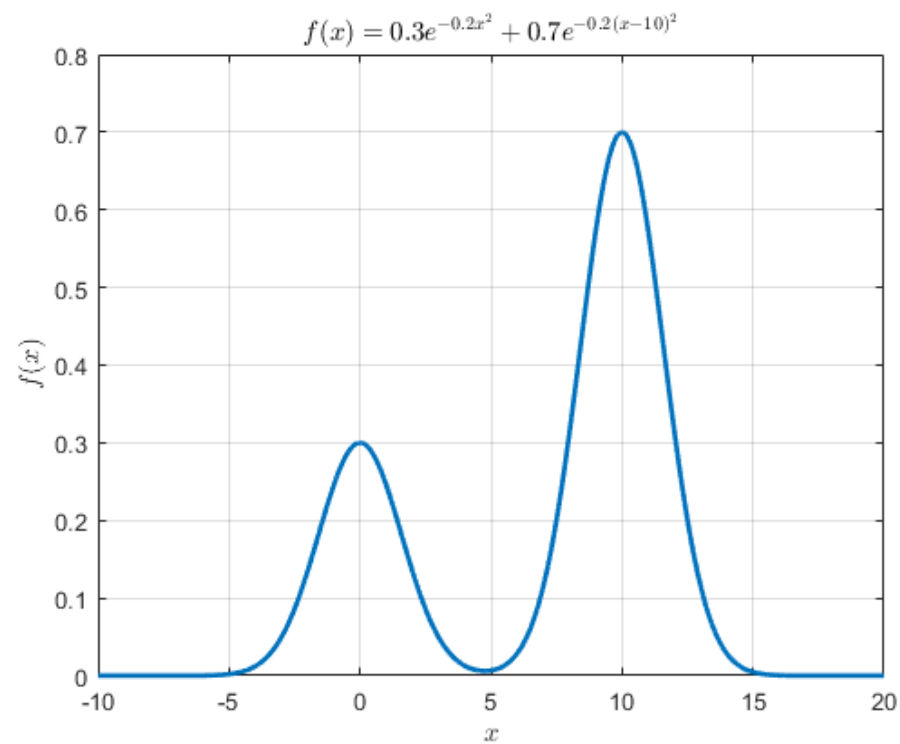
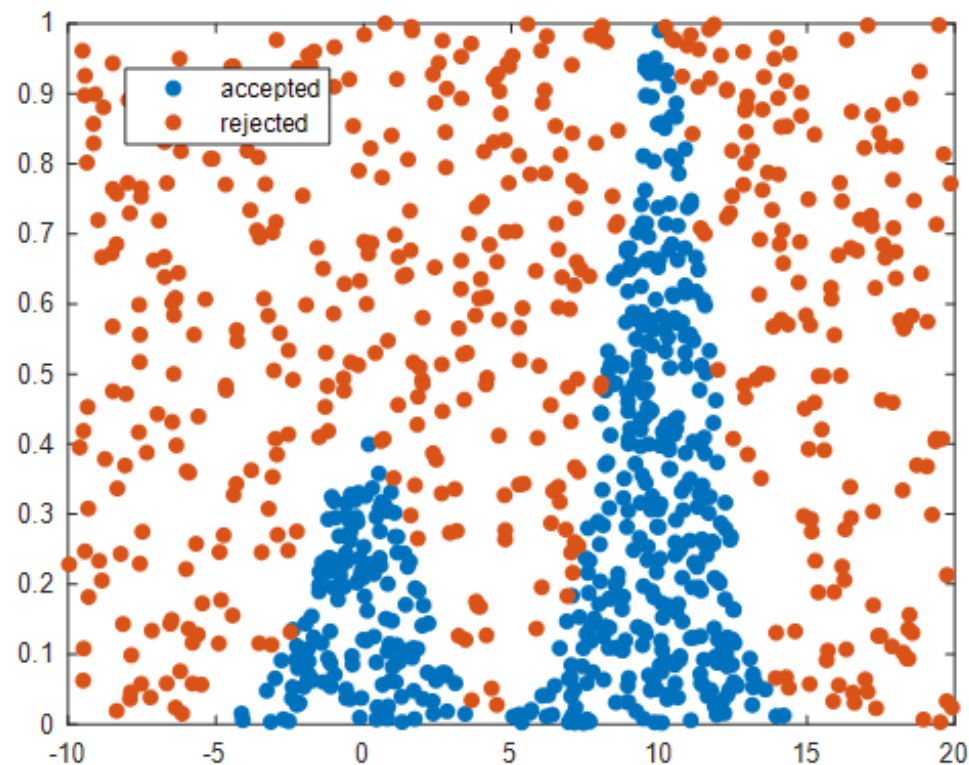
5. Sampling Theory

Rejection sampling



5. Sampling Theory

Rejection sampling



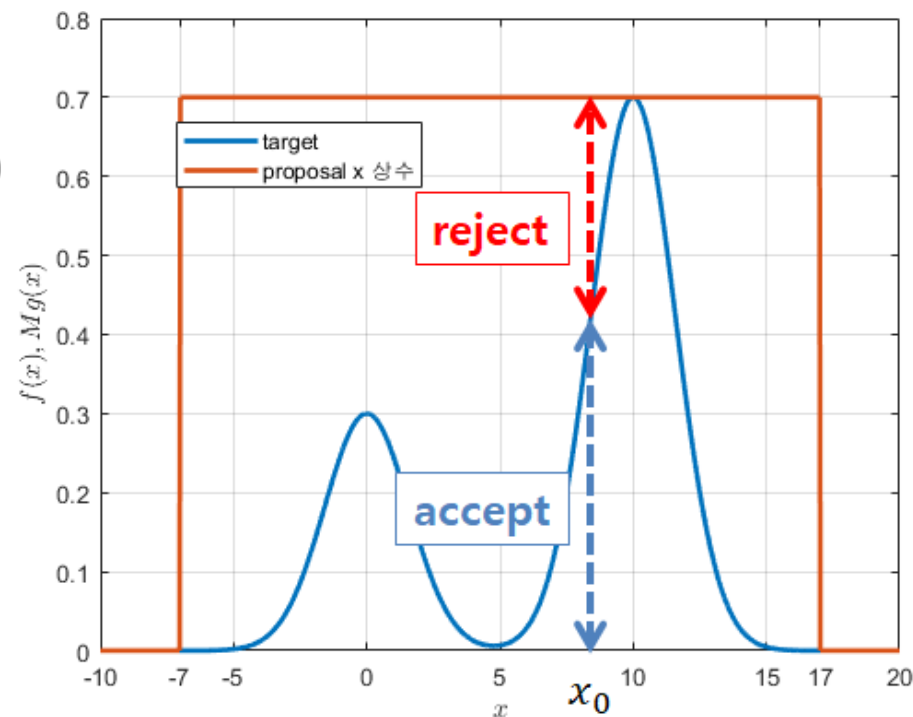
5. Sampling Theory

Rejection sampling

- step 1) **proposal distribution** 설정 $\rightarrow g(x)$
($g(x)$ 는 샘플 생성이 쉬운 Uniform/Normal 분포)
- step 2) generate sample from $g(x) \rightarrow x_0$
& generate random number from $U(0,1)$
- step 3) $U < \frac{f(x_0)}{Mg(x_0)}$ 이면 x_0 을 샘플에 포함

\rightarrow step 2와 step 3을 반복하여 샘플 집단 생성

- 프로젝트에서는
우리가 가진 데이터가 샘플 생성이 쉬우니 $g(x)$ 가 되고
다양한 속성을 가정할 수 있는 정규분포가 **Target distribution**이 된다.
 \rightarrow 우리가 가진 데이터가 정규분포가 아님에도, Rejection sampling으로 뽑은 샘플은 정규분포가 된다!
- 한계점 : reject 되어 버려지는 sample이 많다 \rightarrow Importance sampling, Adaptive rejection sampling ...



5. Sampling Theory

Reservoir sampling

- 언제 사용 ? : 모집단의 크기가 알 수 없을 정도로 커서 한 번에 샘플을 뽑기 어렵고
모집단을 정렬할 기준이 있을 때,
모집단으로부터 하나씩 뽑아 샘플에 넣는 방식으로 샘플링을 하는 경우
ex. 오늘 롯데월드 방문한 사람들 중 500명 (들어온 시간 순서대로 정렬)
→ 모집단에서 하나씩 순서대로 샘플에 들어오는데,
어떻게 뽑아야 “공정하게(무작위로)” 뽑았다고 말할 수 있을까?

5. Sampling Theory

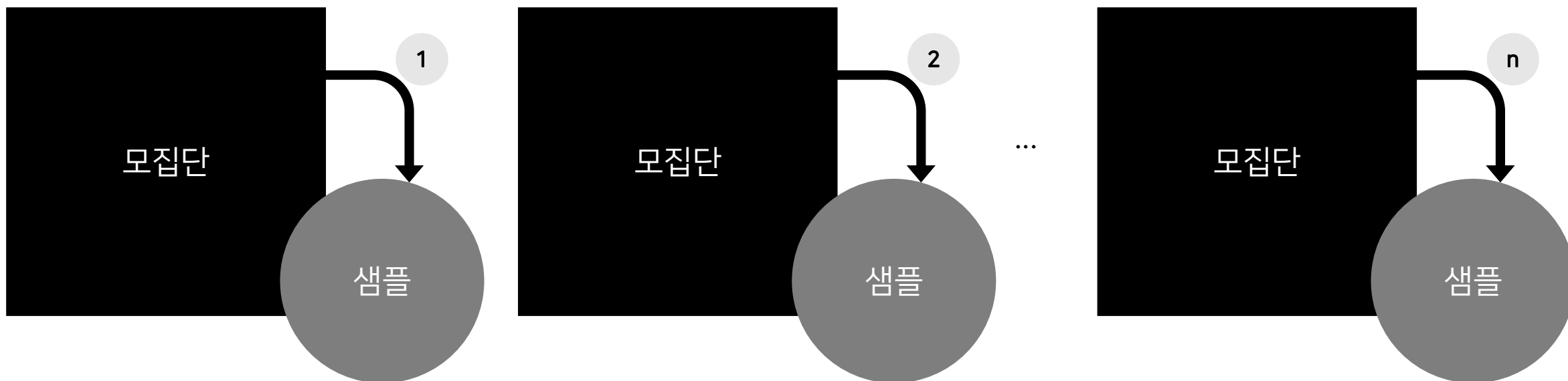
Reservoir sampling

- 언제 사용 ? : 모집단의 크기가 알 수 없을 정도로 커서 한 번에 샘플을 뽑기 어렵고
모집단을 정렬할 기준이 있을 때,
모집단으로부터 하나씩 뽑아 샘플에 넣는 방식으로 샘플링을 하는 경우
ex. 오늘 롯데월드 방문한 사람들 중 500명 (들어온 시간 순서대로 정렬)
→ 모집단에서 하나씩 순서대로 샘플에 들어오는데,
어떻게 뽑아야 “공정하게(무작위로)” 뽑았다고 말할 수 있을까?
- 공정하다 = 모집단의 모든 대상들에 대해 샘플에 포함될 확률이 각각 동일하다
(10명 중 3명을 무작위로 뽑을 때, 10명 각자의 입장에서는 ‘내가 뽑힐 확률’이 3/10로 동일)
그러므로 N명 중에서 n명을 뽑는다면 N명 모두가 ‘내가 뽑힐 확률’이 n/N 이어야 공정하다

5. Sampling Theory

Reservoir sampling

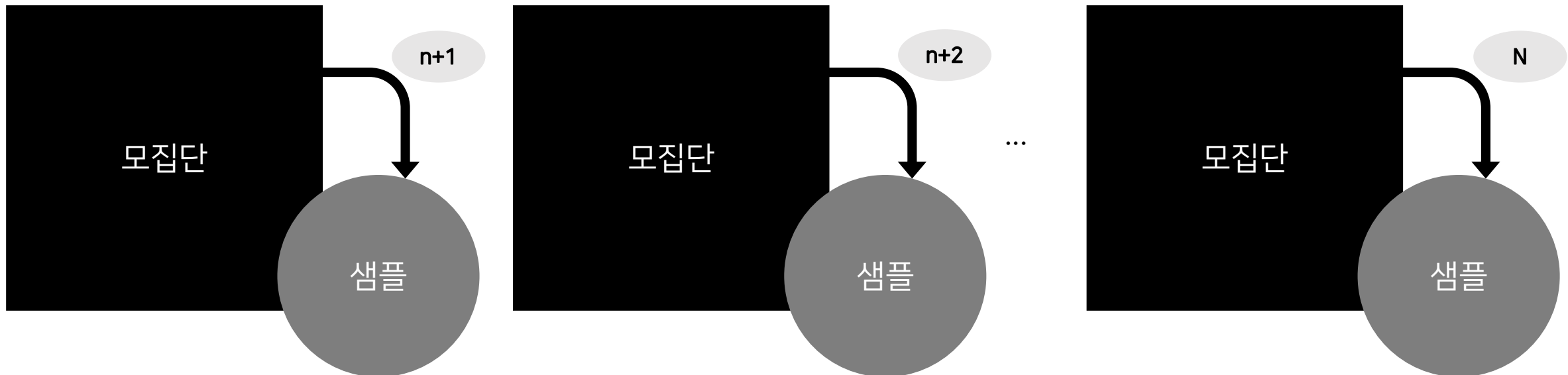
- 모집단으로부터 n 개의 샘플을 뽑는다고 할 때,
첫 번째로 뽑힌 대상은 샘플에 넣고, 두 번째, 세 번째, \dots , n 번째로 뽑힌 대상까지는 일단 샘플에 넣는다



5. Sampling Theory

Reservoir sampling

- 샘플이 n 개가 되면,
모집단으로부터 뽑힌 $n+1$ 번째 대상이 샘플에 들어갈지는 $\frac{n}{n+1}$ 확률로 결정하고, 만약 들어가게 되면
샘플에 있는 n 개의 대상들 중 하나를 $1/n$ 확률로 뽑아 샘플에서 제외한다



5. Sampling Theory

Reservoir sampling 예시

- 10개 중 3개를 뽑는다면,
- 모집단으로부터 1~3번째로 온 대상 → 일단은 샘플에 들어가고 → 끝까지 제외되지 않으면 → 샘플로 뽑힌 것

$$P = 1 * \left(1 - \frac{3}{4} * \frac{1}{3}\right) * \left(1 - \frac{3}{5} * \frac{1}{3}\right) * \dots * \left(1 - \frac{3}{10} * \frac{1}{3}\right)$$
$$= 1 * \frac{3}{4} * \frac{4}{5} * \dots * \frac{9}{10} = \frac{3}{10}$$

- 모집단으로부터 4~10번째로 온 대상 → 확률에 따라 샘플에 들어가고 → 끝까지 제외되지 않으면 → 샘플로 뽑힌 것

$$P = \frac{3}{4} * \left(1 - \frac{3}{5} * \frac{1}{3}\right) * \dots * \left(1 - \frac{3}{10} * \frac{1}{3}\right)$$
$$= \frac{3}{4} * \frac{4}{5} * \dots * \frac{9}{10} = \frac{3}{10}$$

“N명 중에서 n명을 뽑는다면 N명 모두가 내가 뽑힐 확률이 n/N이어야 공정하다”

5. Sampling Theory

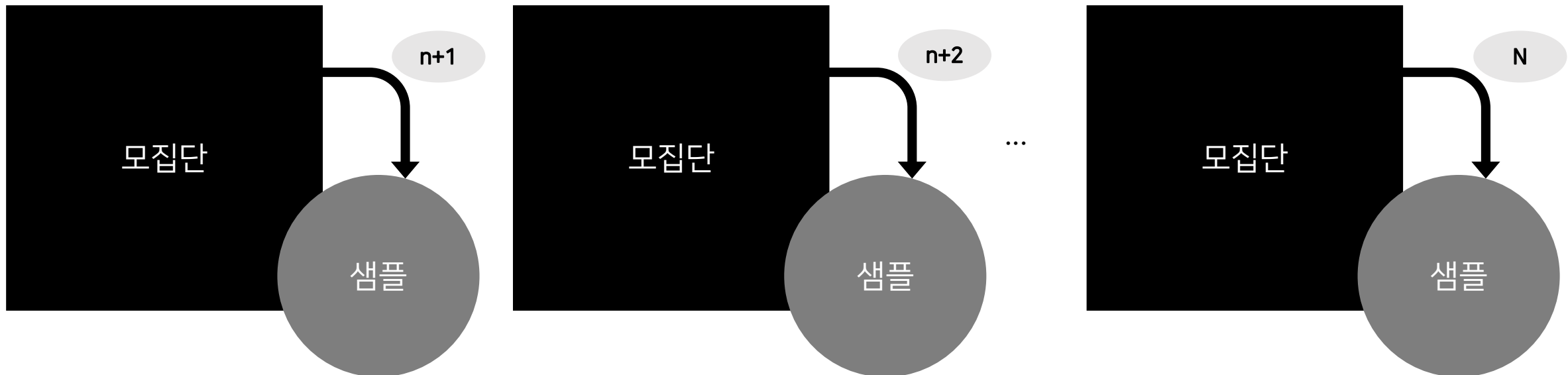
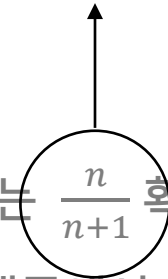
Reservoir sampling

- 샘플이 n 개가 되면,

모집단으로부터 뽑힌 $n+1$ 번째 대상이 샘플에 들어갈지는 $\frac{n}{n+1}$ 확률로 결정하고, 만약 들어가게 되면

샘플에 있는 n 개의 대상들 중 하나를 $1/n$ 확률로 뽑아 샘플에서 제외한다

0에 수렴하므로 모집단의 모든 대상을 거치지 않아도
sample은 변하지 않는 상태로 수렴



Summary

- 왜 통계를 공부해야 할까? → 숫자로 의사소통 하는 방법, 문제에 적절한 해결방법을 사용하기 위해
- t-test : 두 표본의 평균의 차이와 표준편차의 비율을 통해 평균의 차이가 유의미한지 검정
- ANOVA : 3개 이상의 표본 간 분산과 표본 내 분산을 비교하여 적어도 하나의 표본의 평균이 다른지 검정
- Regression : 독립변수로 종속변수를 예측하는 통계적 방법론. t-test로 추세선의 유의성 검정
- Sampling : 샘플을 한 번 뽑을 때 모집단의 특성이 잘 반영되도록, or 우리가 다루기 쉬운 분포를 따르도록
ex. Rejection sampling, Reservoir sampling ...

6. Summary

Reference

- Youtube "Sapientia a Dei"
- Youtube " StatQuest with Josh Starmer"
- 고급데이터사이언스방법론 (임종호 교수님) 강의안
- Blog 공돌이의 수학정리노트

DATA SCIENCE LAB

발표자 : 정건우 010-6473-3938
E-mail : wjdrjsdn3938@naver.com