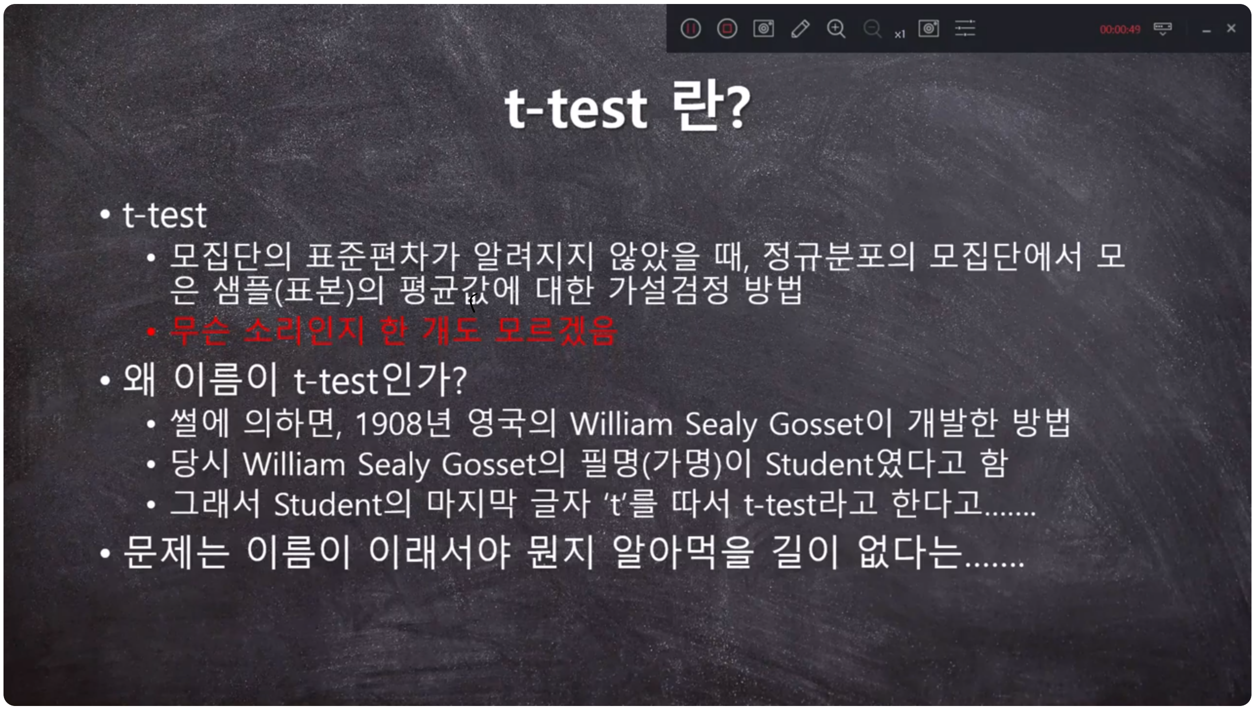
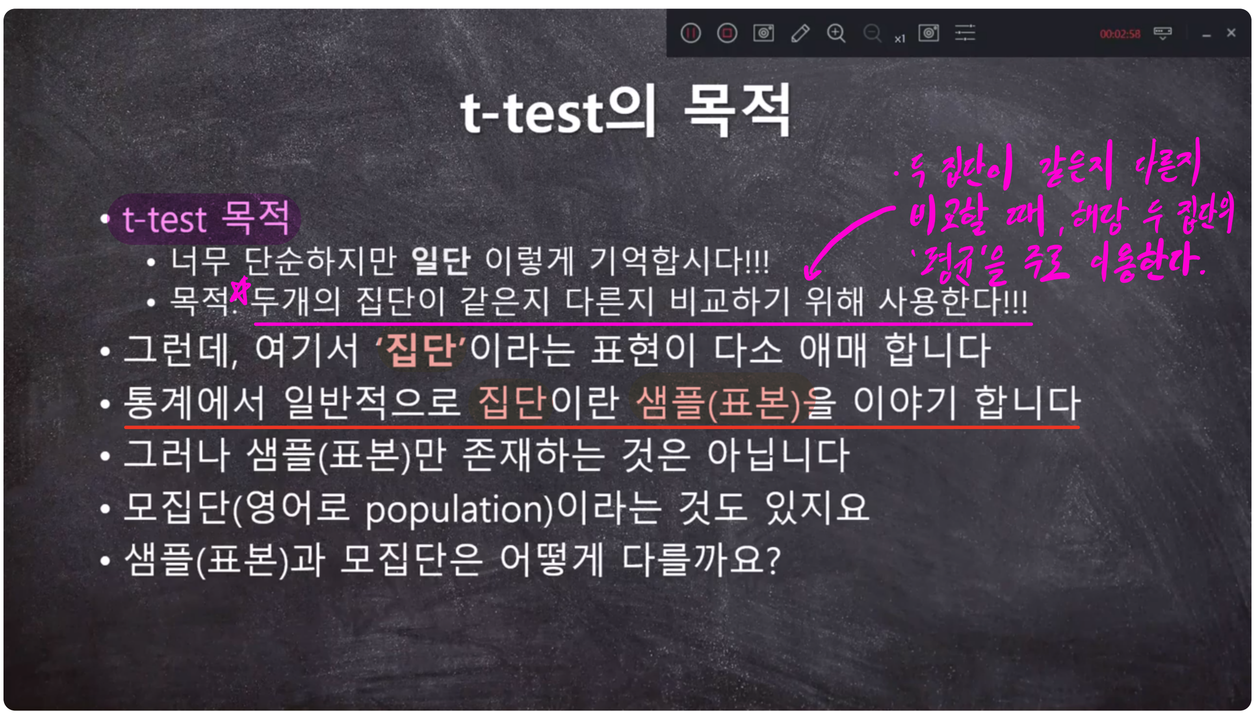
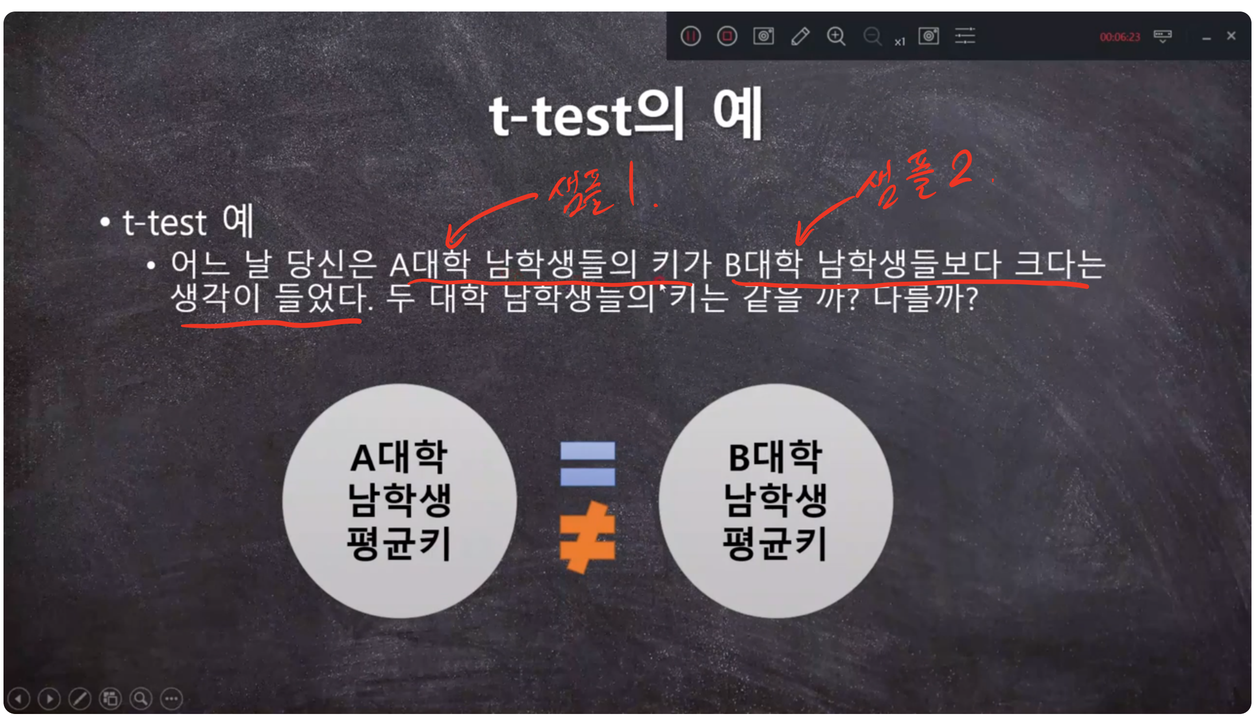




※ '두 집단'은 사실 '두 모집단'을 의미한다. 실제로 두 모집단을 조사하여 비교하는 것은 매우 어렵기 때문에, '두 표본'을 추출하여, 해당 두 표본을 비교하는 것이다. 결국, '표본'을 이용하여 '모집단'을 추정하는 맥락에 속한다.

<거의 모든 통계 분석 상황에선 '통계량'을 활용한다.>



t-test의 예

- t-test 예
  - A대학 남학생 평균키 = 178.5cm
  - B대학 남학생 평균키 = 179.9cm

A대학
178.5cm

= ≠

B대학
179.9cm

t-test의 예

A대학
178.5cm

= ≠

B대학
179.9cm

<우연히 발생했다
=큰 의미를 둘 수 없다.>

- t-test를 위한 통계적 질문
  - A대학 남학생 평균키(178.5cm)와 B대학 남학생 평균키(179.9cm)가 우연히 같은 확률은 얼마나 될까?
  - A대학과 B대학의 남학생 평균키 차이인 1.4cm가 우연히 발생했을 확률은 얼마나 될까?

같은 질문임

↳ 평균 키차이인 1.4cm가
우연히 발생했다면,
키 차이가 존재하지 않음.
=) 우연히 발생한게 아니라면,

(의미 있는 키 차이가 존재함
(유의한, 인과관계가 있는)

t-test에 대한 보다 깊은 이해

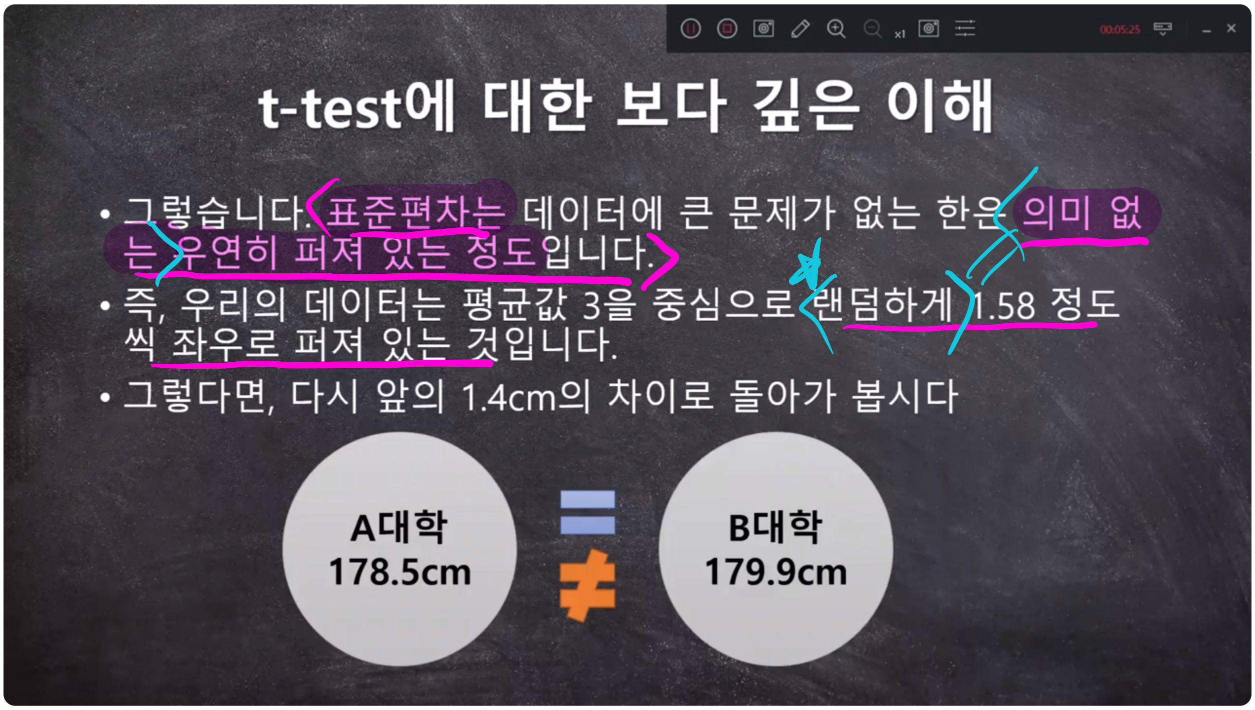
A대학 178.5cm ≠ B대학 179.9cm

- t-test를 위한 통계적 질문
  - 그렇다면 과연 1.4cm의 차이가 얼마나 커야 우연히 발생하지 않았다고 판단 할 수 있을까?
  - 1.4cm의 차이는 과연 큰 것인가 작은 것인가?

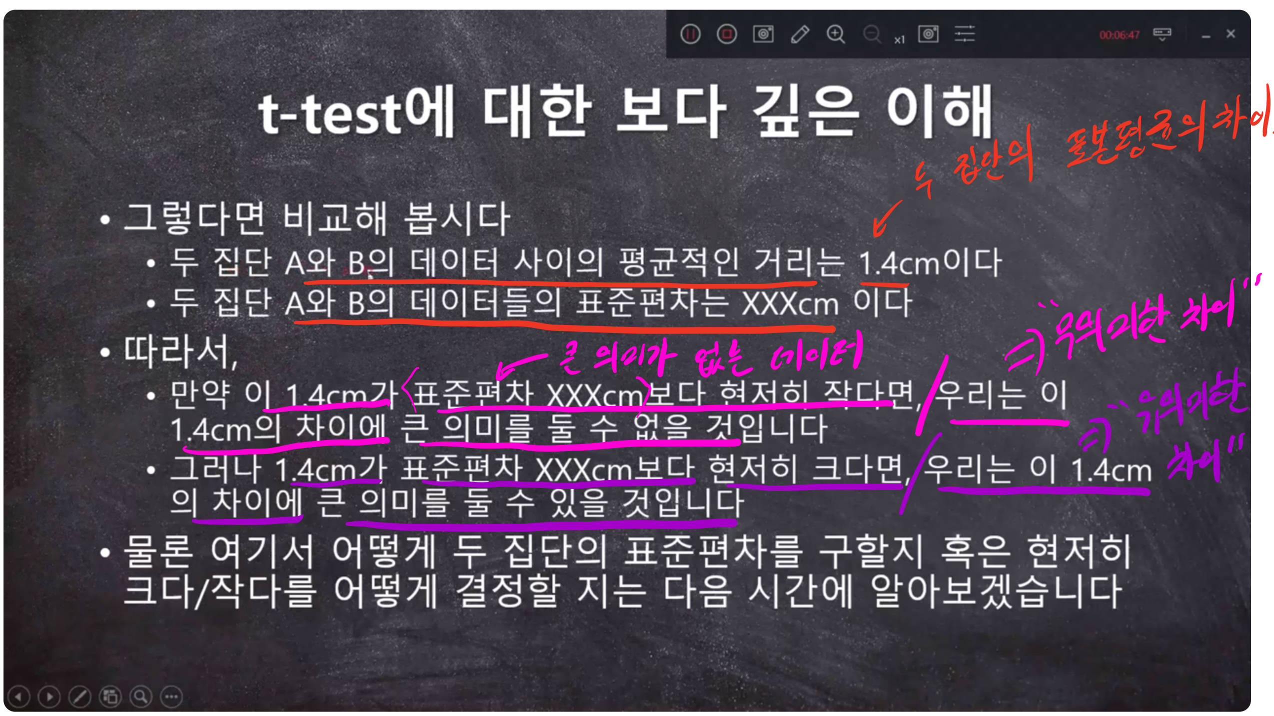
t-test에 대한 보다 깊은 이해

- 우리는 1.4cm가 얼마나 큰지 혹은 작은지 알 수 없습니다!!
- 우리는 이제 이 1.4cm가 얼마나 큰지 혹은 작은지 결정할 나름의 비교 대상이 필요합니다.
- 누구를 가지고 와서 비교해야 할까요?

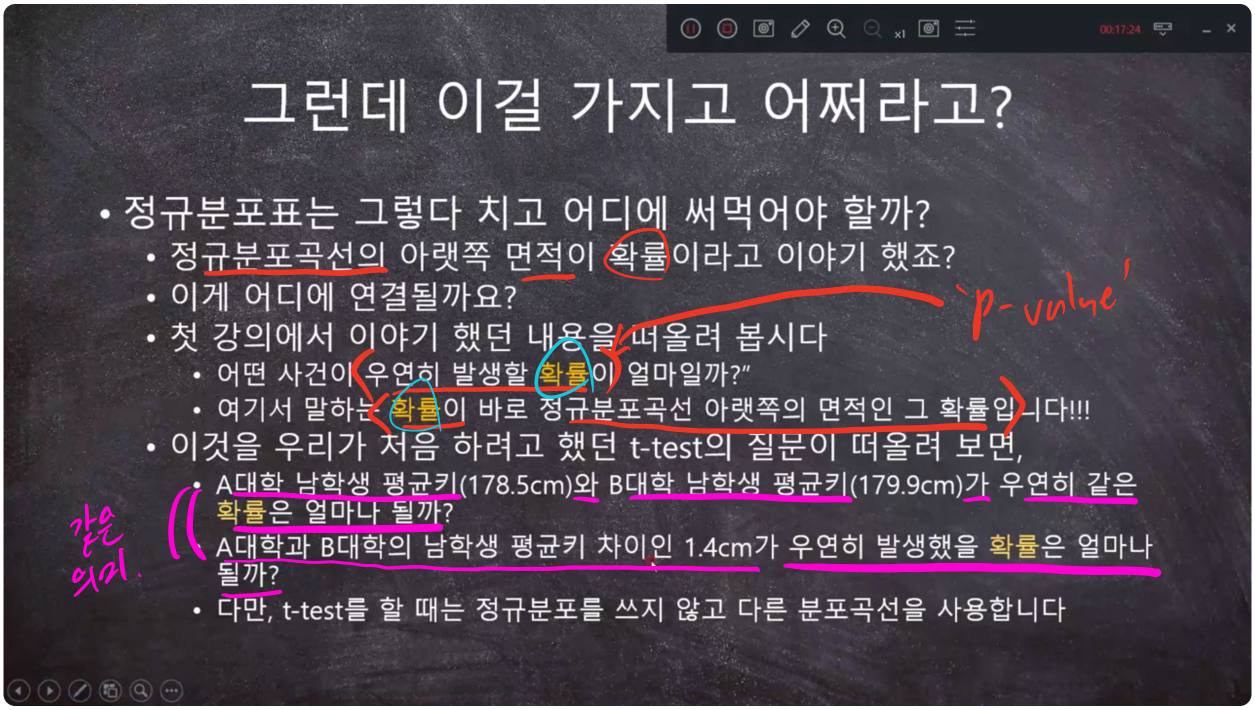
표준편차(분산)

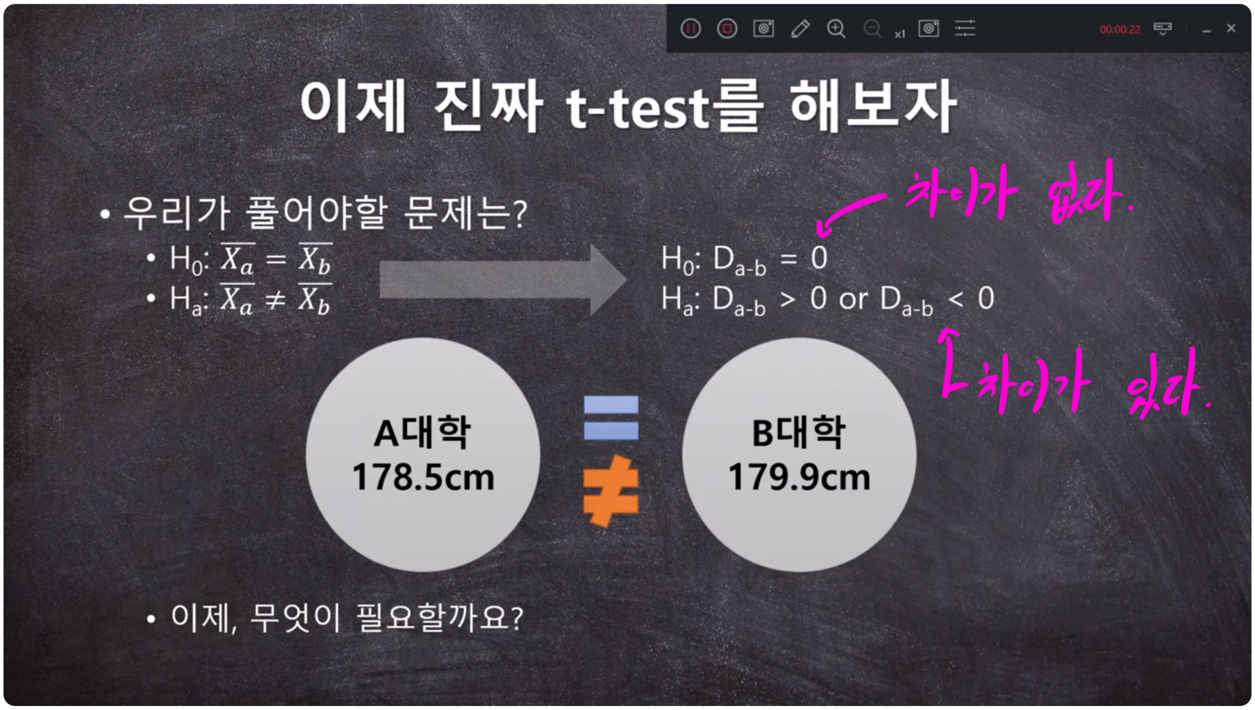


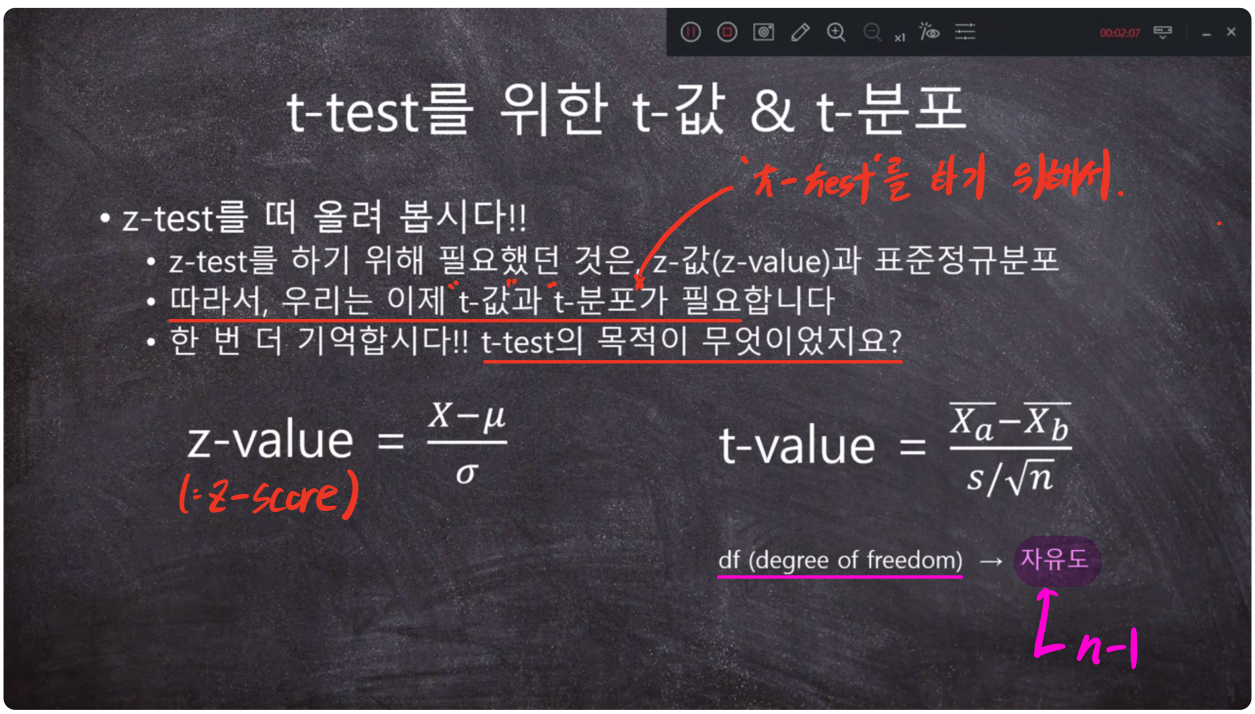
t-test에 대한 보다 깊은 이해
• 그렇습니다. 표준편차는 데이터에 큰 문제가 없는 한은 의미 없는 우연히 퍼져 있는 정도입니다.
• 즉, 우리의 데이터는 평균값 3을 중심으로 랜덤하게 1.58 정도씩 좌우로 퍼져 있는 것입니다.
• 그렇다면, 다시 앞의 1.4cm의 차이로 돌아가 봅시다
A대학
178.5cm
B대학
179.9cm



t-test에 대한 보다 깊은 이해
두 집단의 표본평균의차이.
• 그렇다면 비교해 봅시다
• 두 집단 A와 B의 데이터 사이의 평균적인 거리는 1.4cm이다
• 두 집단 A와 B의 데이터들의 표준편차는 XXXcm 이다
• 따라서,
큰 의미가 없는 데이터
=) "유의미한 차이"
• 만약 이 1.4cm가 표준편차 XXXcm보다 현저히 작다면, 우리는 이 1.4cm의 차이에 큰 의미를 둘 수 없을 것입니다
=) "유의미한 차이"
• 그러나 1.4cm가 표준편차 XXXcm보다 현저히 크다면, 우리는 이 1.4cm의 차이에 큰 의미를 둘 수 있을 것입니다
• 물론 여기서 어떻게 두 집단의 표준편차를 구할지 혹은 현저히 크다/작다를 어떻게 결정할 지는 다음 시간에 알아보겠습니다


그런데 이걸 가지고 어쩌라고?
• 정규분포표는 그렇다 치고 어디에 써먹어야 할까?
• 정규분포곡선의 아랫쪽 면적이 확률이라고 이야기 했죠?
• 이게 어디에 연결될까요?
• 첫 강의에서 이야기 했던 내용을 떠올려 봅시다
• 어떤 사건이 우연히 발생할 확률이 얼마일까?"
• 여기서 말하는 확률이 바로 정규분포곡선 아랫쪽의 면적인 그 확률입니다!!!
'p-value'
• 이것을 우리가 처음 하려고 했던 t-test의 질문이 떠올려 보면,
• A대학 남학생 평균키(178.5cm)와 B대학 남학생 평균키(179.9cm)가 우연히 같은 확률은 얼마나 될까?
• A대학과 B대학의 남학생 평균키 차이인 1.4cm가 우연히 발생했을 확률은 얼마나 될까?
같은 의미.
• 다만, t-test를 할 때는 정규분포를 쓰지 않고 다른 분포곡선을 사용합니다


이제 진짜 t-test를 해보자
• 우리가 풀어야할 문제는?
• H_0: $\overline{X_a} = \overline{X_b}$
• H_a: $\overline{X_a} \neq \overline{X_b}$
차이가 없다.
H_0: $D_{a\text{-}b} = 0$
H_a: $D_{a\text{-}b} > 0$ or $D_{a\text{-}b} < 0$
차이가 있다.
A대학
178.5cm
B대학
179.9cm
• 이제, 무엇이 필요할까요?


t-test를 위한 t-값 & t-분포
't-test'를 하기 위해서.
• z-test를 떠 올려 봅시다!!
• z-test를 하기 위해 필요했던 것은, z-값(z-value)과 표준정규분포
• 따라서, 우리는 이제 "t-값"과 "t-분포"가 필요합니다
• 한 번 더 기억합시다!! t-test의 목적이 무엇이었지요?
z-value $= \frac{X-\mu}{\sigma}$
(=z-score)
t-value $= \frac{\overline{X_a}-\overline{X_b}}{s/\sqrt{n}}$
df (degree of freedom) → 자유도
n-1

t-값 (t-value)의 의미

- 자, 다시 들여다 봅시다

$$\text{t-value} = \frac{\overline{X_a}-\overline{X_b}}{s/\sqrt{n}}$$

- 우리의 목적은 두 집단의 평균값이 같은지 다른지 알고 싶습니다
- 그래서 통계적 가설에 의거하여 두 집단의 평균값의 차이가 "0"과 같은 지 다른 지 궁금합니다
- 위의 값에서 우리가 궁금해 하는 그 차이는 분자에 있습니다
- 여기서부터 중요한 것이 통계적인 생각/질문/접근법입니다
- 도대체 "저 값이" 얼마나 커야 큰 것일까요?
- 비교 대상이 필요하겠죠?

"두 표본평균의 차이가"

t-값 (t-value)의 의미

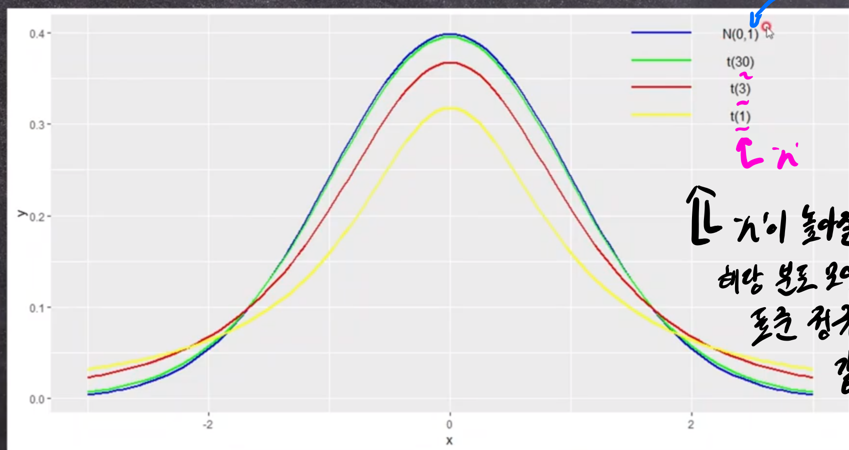
- 그렇다면

$$\text{t-value} = \frac{\overline{X_a}-\overline{X_b}}{s/\sqrt{n}} = \frac{\overline{X_a}-\overline{X_b}}{\frac{s}{\sqrt{n}}}$$

표준오차

- 우리는 이 두 평균값의 차이를 표준편차와 비교하는 겁니다
- 왜?
  - 표준편차란 우리의 데이터가 평균값을 기준으로 평균적으로 퍼진 정도입니다
  - 따라서 이 자체는 의미 없는 편차
  - 만약 두 집단의 평균값의 (편)차가 의미 없는 편차인 표준편차만도 못하다면 당연히, 이 차이는 우연히 발생했다고 보아야 할 것입니다
  - 그런데, 여기서 $\sqrt{n}$의 역할은 무엇일까요?

$\sqrt{n}$의 역할



표준정규분포

n

n이 높아질수록
해당 분포 모양이
표준 정규 분포 모양과
같게 됨.

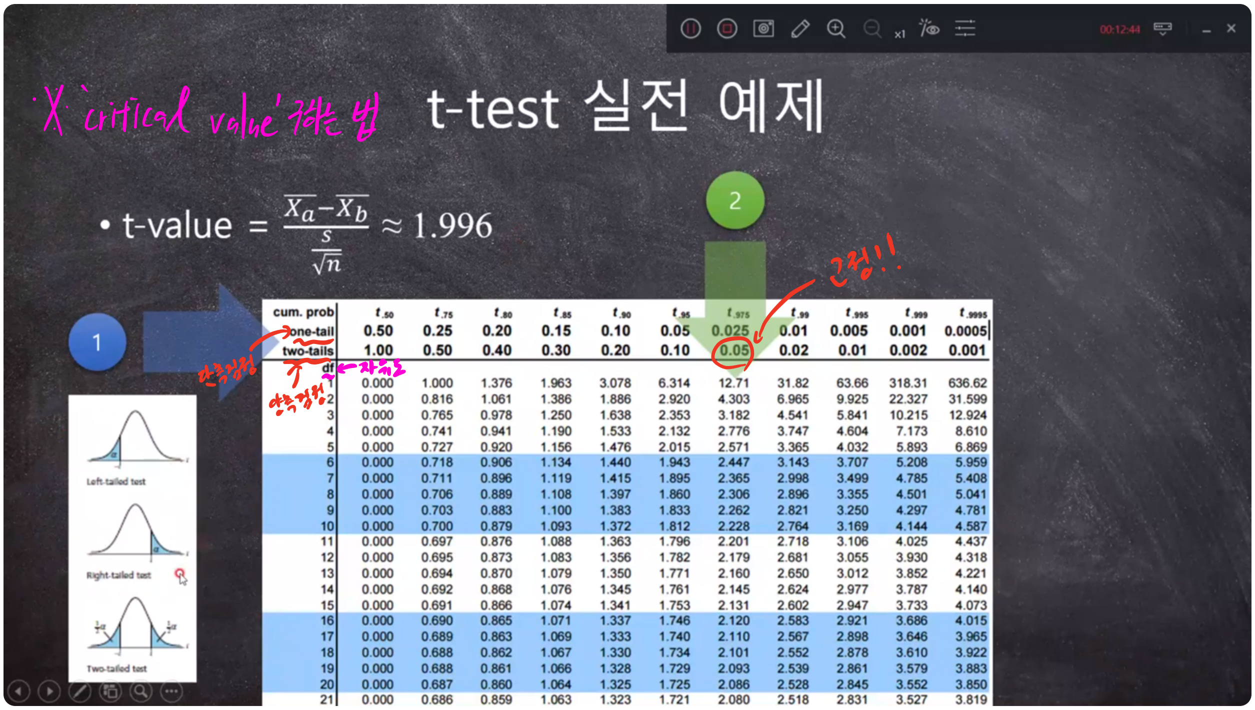
t-값 (t-value)의 의미

- 정리하자면

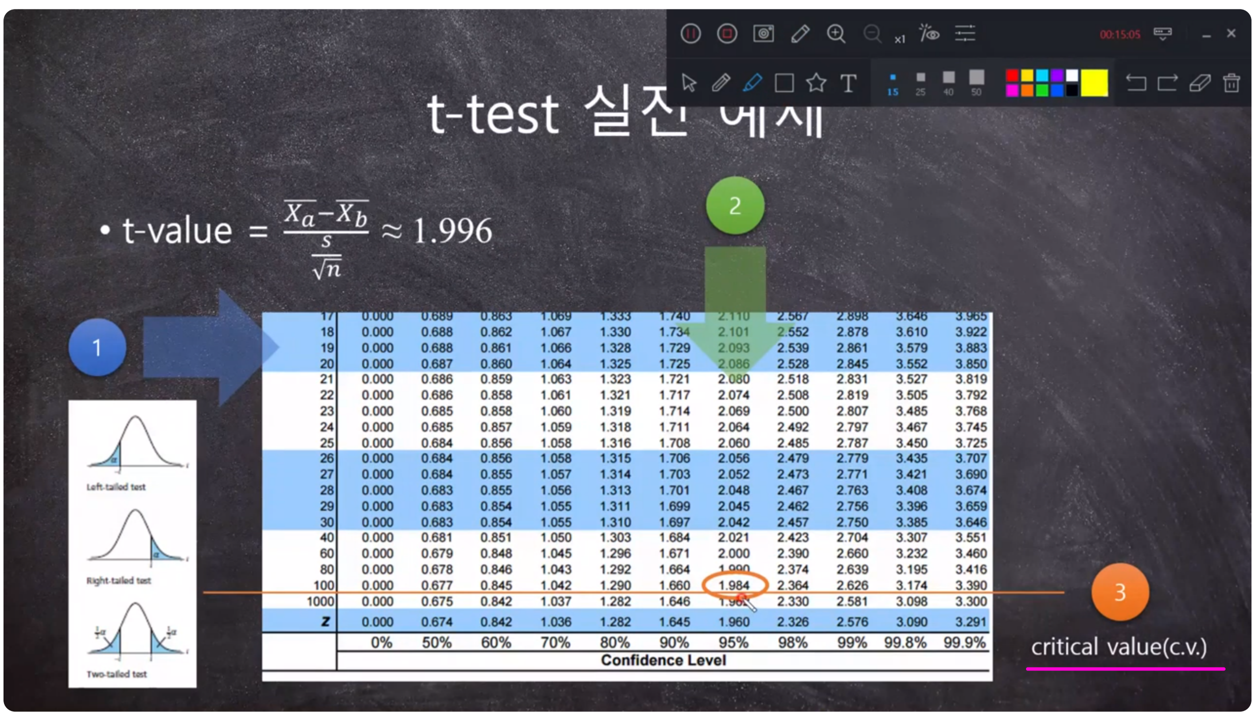
$$\text{t-value} = \frac{\overline{X_a} - \overline{X_b}}{s/\sqrt{n}} = \frac{\overline{X_a} - \overline{X_b}}{\frac{s}{\sqrt{n}}}$$

  - 두 집단의 평균값의 (편)차이가 의미 없는 편차인 표준편차 만도 못하다면, 이 차이는 우연히 발생했다고 본다
  - ① 표본의 크기 (n)이 커지면 커질 수록 t-값은 커지고
  - 표본의 크기 (n)이 커지면 커질 수록, t-분포는 표준 정규분포에 근사
  - ② t-test에서 자유도(df)는 n-1로 계산되므로, 표본의 크기가 커지면 자유도가 커지고, 자유도가 커졌다는 의미는 우리가 t-분포에 묶여 있다가 자유롭게 표준정규분포를 사용할 수 있음을 의미

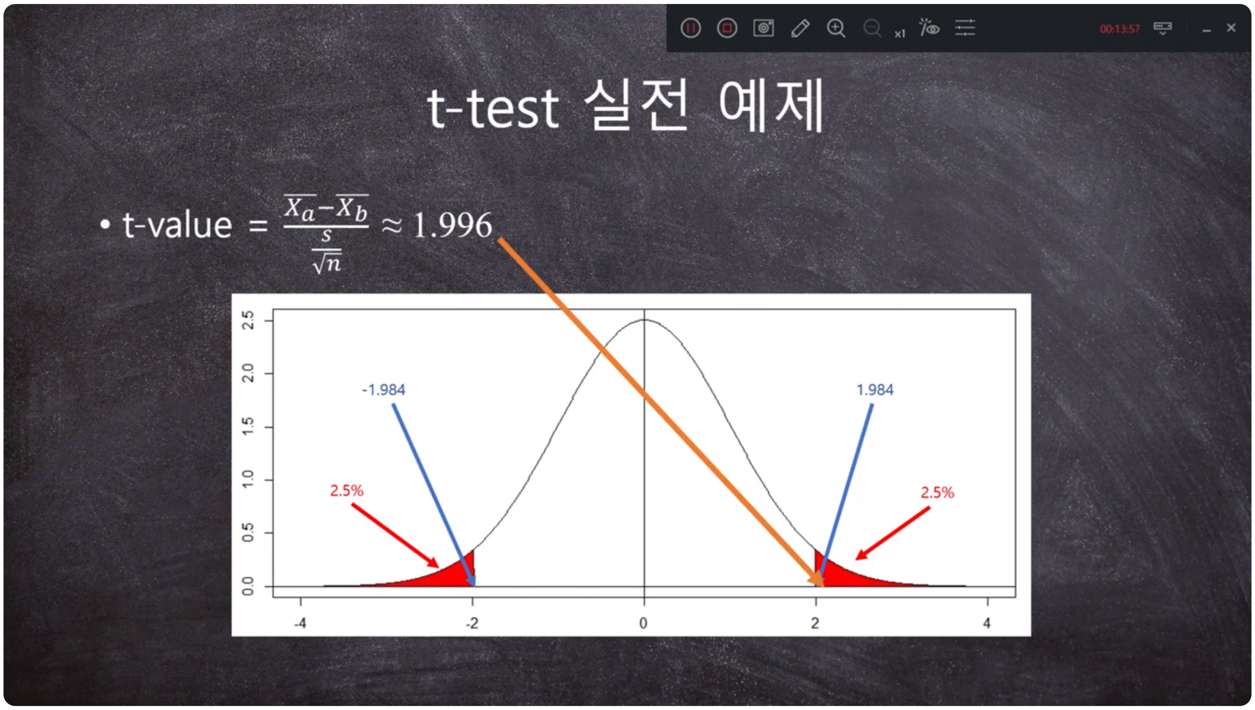
※ 'critical value' 구하는 법

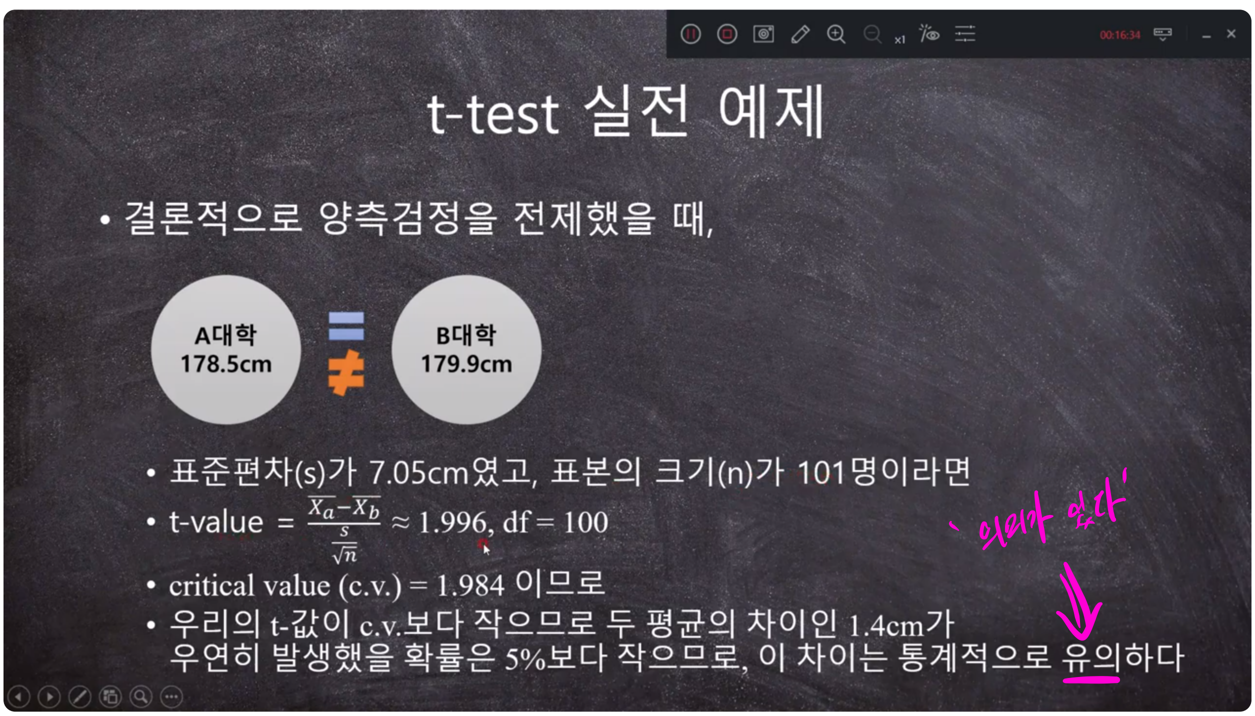


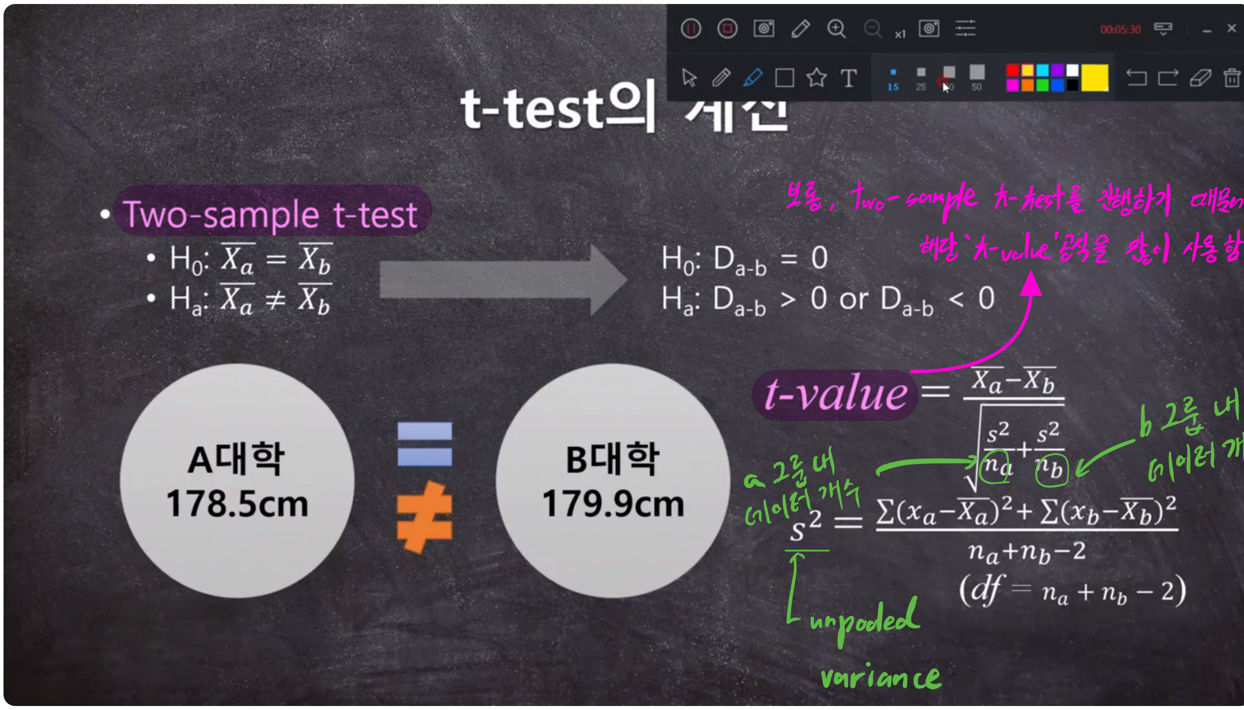
cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819

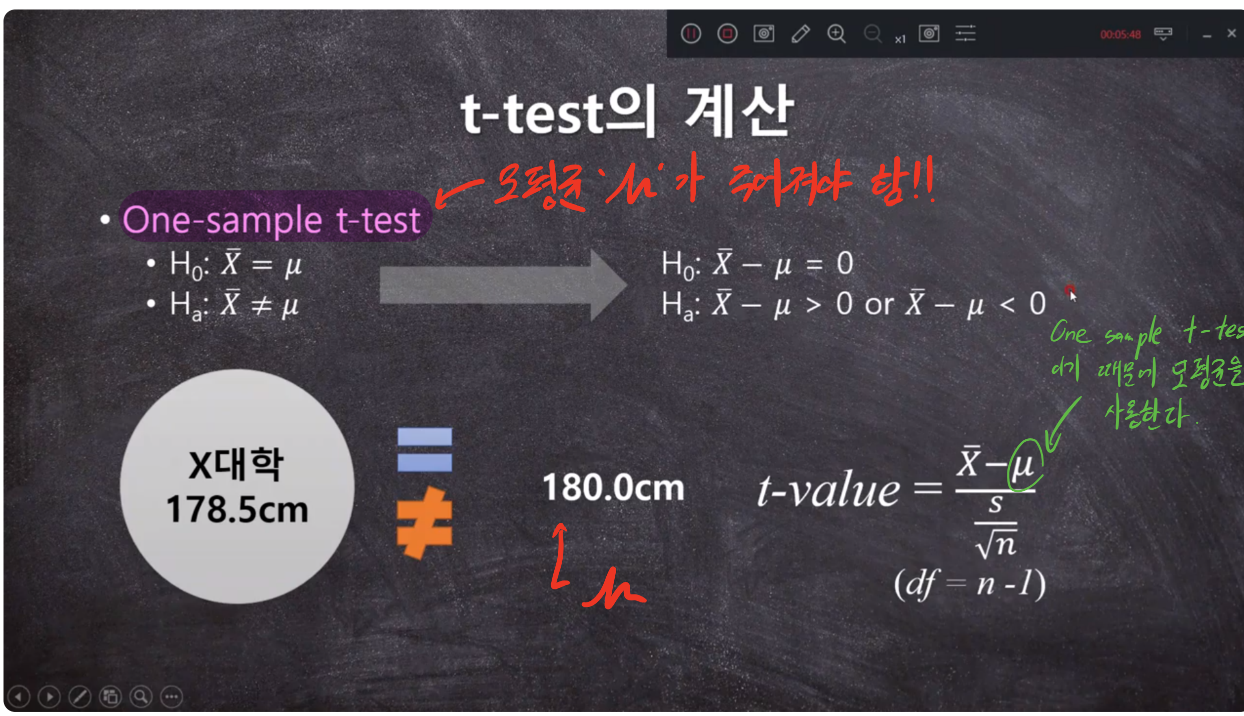


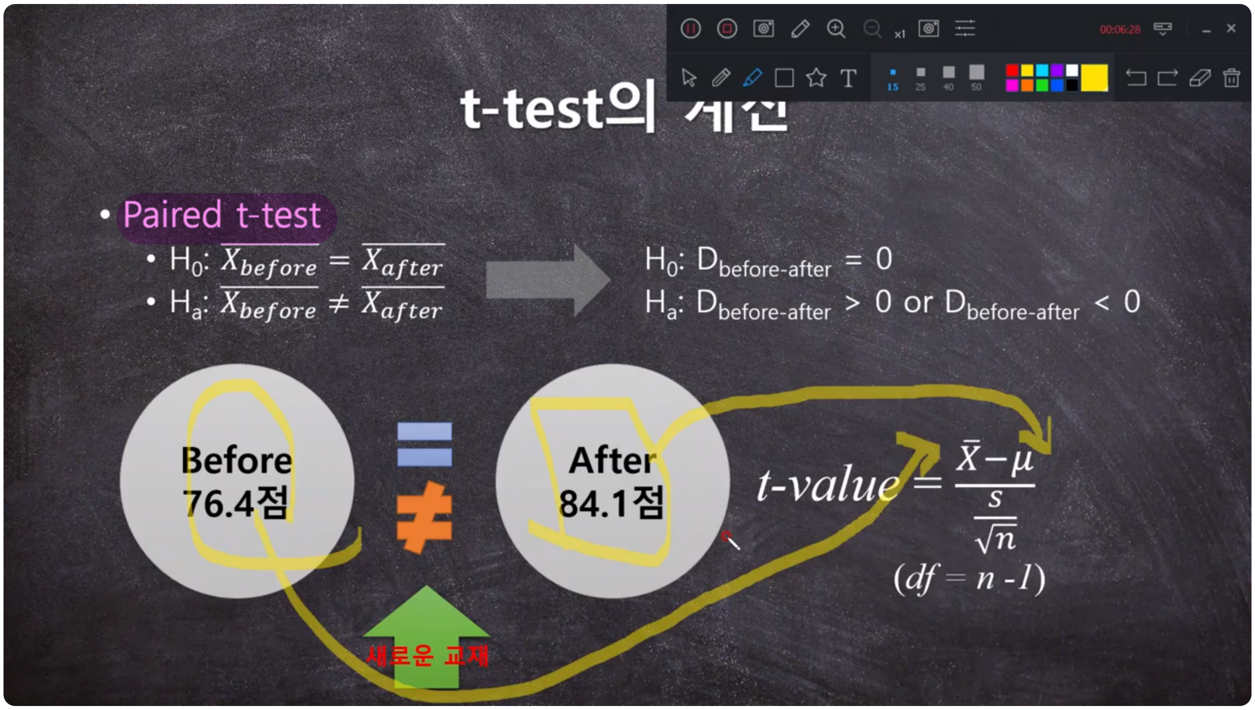
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	[illegible]	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										


t-test 실전 예제
• t-value = $\frac{\overline{X_a} - \overline{X_b}}{\frac{s}{\sqrt{n}}} \approx 1.996$
-1.984
1.984
2.5%
2.5%


t-test 실전 예제
• 결론적으로 양측검정을 전제했을 때,
A대학 178.5cm
B대학 179.9cm
• 표준편차(s)가 7.05cm였고, 표본의 크기(n)가 101명이라면
• t-value = $\frac{\overline{X_a} - \overline{X_b}}{\frac{s}{\sqrt{n}}} \approx 1.996$, df = 100
• critical value (c.v.) = 1.984 이므로
• 우리의 t-값이 c.v.보다 작으므로 두 평균의 차이인 1.4cm가
우연히 발생했을 확률은 5%보다 작으므로, 이 차이는 통계적으로 유의하다
'의미가 있다'


t-test의
Two-sample t-test
H_0: $\overline{X_a} = \overline{X_b}$
H_a: $\overline{X_a} \neq \overline{X_b}$
H_0: $D_{a-b} = 0$
H_a: $D_{a-b} > 0$ or $D_{a-b} < 0$
보통, two-sample t-test를 진행하기 때문에
해당 't-value' 공식을 많이 사용함.
A대학
178.5cm
B대학
179.9cm
$t\text{-value} = \dfrac{\overline{X_a}-\overline{X_b}}{\sqrt{\dfrac{s^2}{n_a}+\dfrac{s^2}{n_b}}}$
a 그룹 내 데이터 개수
b 그룹 내 데이터 개수.
$s^2 = \dfrac{\sum(x_a-\overline{X_a})^2+\sum(x_b-\overline{X_b})^2}{n_a+n_b-2}$
$(df = n_a + n_b - 2)$
unpooled variance


t-test의 계산
One-sample t-test
모평균 'μ'가 주어져야 함!!
H_0: $\bar{X} = \mu$
H_a: $\bar{X} \neq \mu$
H_0: $\bar{X} - \mu = 0$
H_a: $\bar{X} - \mu > 0$ or $\bar{X} - \mu < 0$
One sample t-test 이기 때문에 모평균을 사용한다.
X대학
178.5cm
180.0cm
μ
$t\text{-value} = \dfrac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$
$(df = n - 1)$



이 결과를 논문 형태 표로 정리하면, 대략 아래 표와 같습니다.

M(Mean): 평균, SD(Standard Deviation): 표준편차 이니 참고하시구요.

브랜드	M	SD	t	P
삼송	3.14	1.32	2.393*	.017
LZ	2.78	1.39		

* p<.05

① 만족도에 대한 두 그룹의 평균차이가 있다.

결과적으로,

② "삼송과 LZ의 만족도는 통계적으로 유의미한 차이가 있으며(p<.05),
③ 삼송 제품 사용자의 휴대폰 만족도가 LZ보다 높은 것으로 판단되었다."

이런 식으로 결론을 내어주면 되겠습니다.

A반의 1학기 평균이 100점 만점에 **30**점이고

2학기 평균이 100점 만점에 **31**점이라고 한다면

두 집단, 즉 1학기와 2학기의 평균 차이는 고작 **1**점입니다.

자세한 계산은 하지 않겠지만, 만점이 100점인 사실로 미루어볼 때

1점의 평균 차이는 **그다지 유의하지 않다**는 결과가 나올 것입니다.

그러나 만약 1학기 평균이 **30**점으로 동일하고

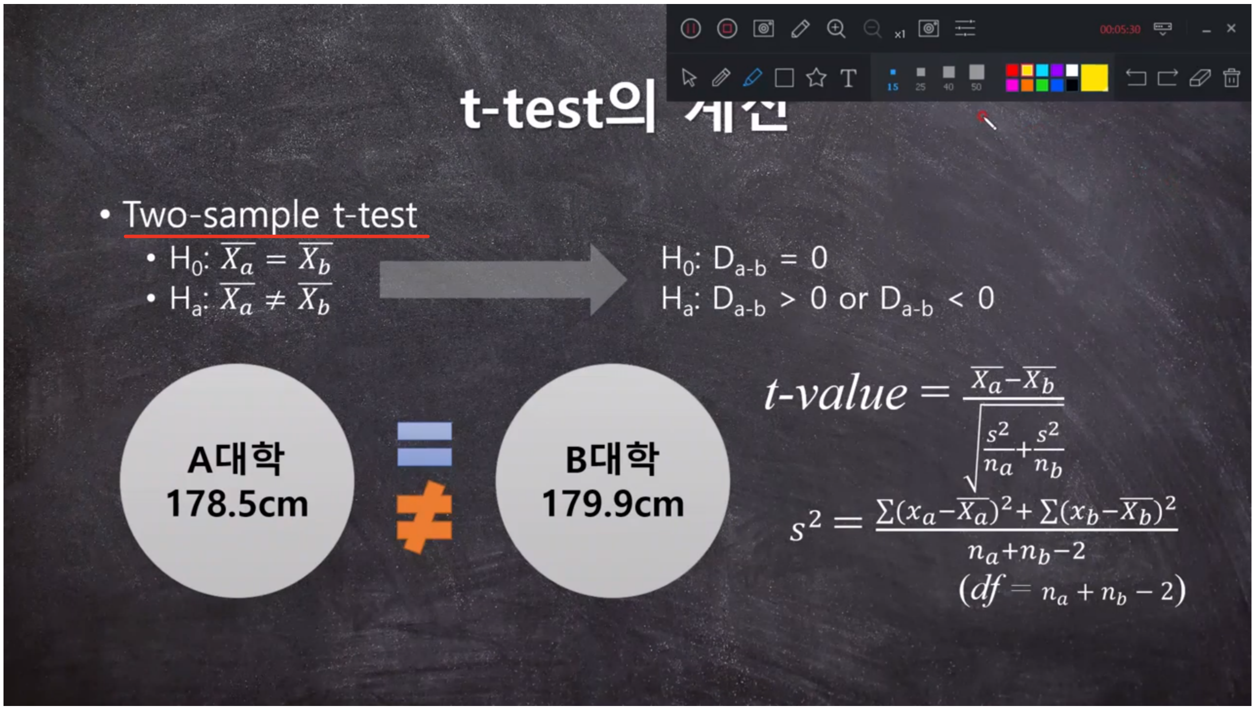
2학기 평균은 **90**점이라면

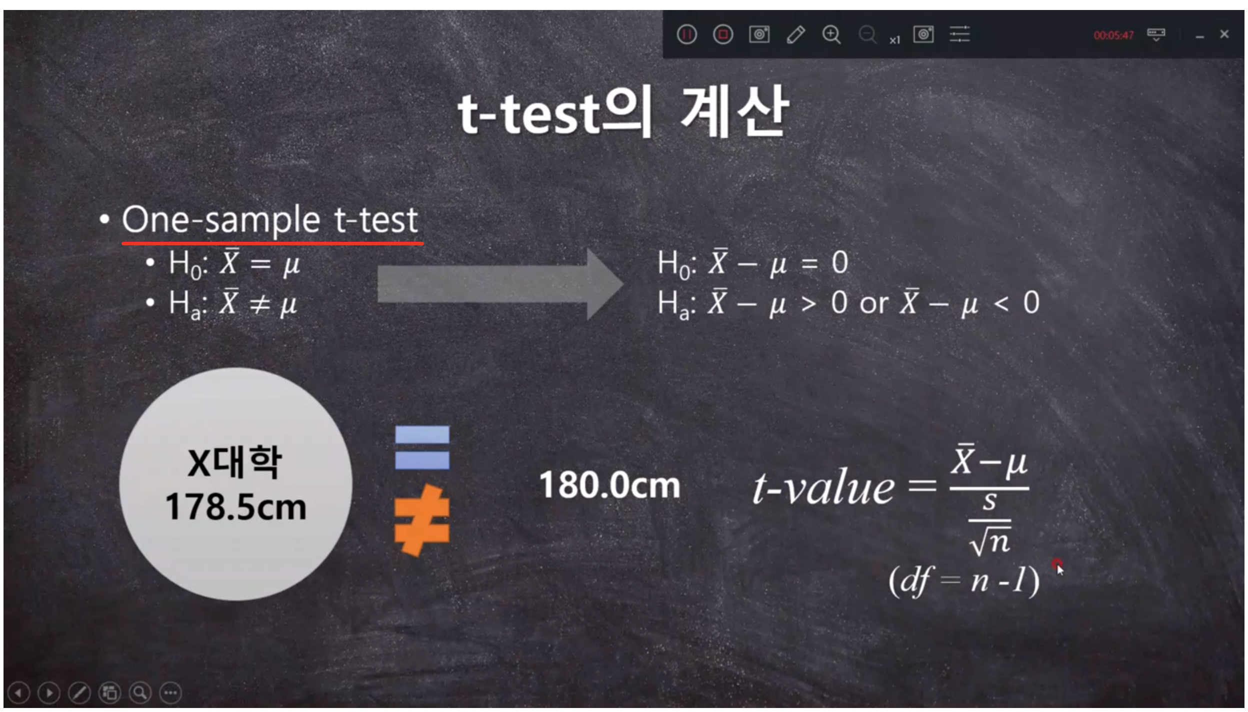
두 집단의 평균 차이는 **60**점으로 꽤 **유의미한 차이**라는 결과가 나타나겠죠?

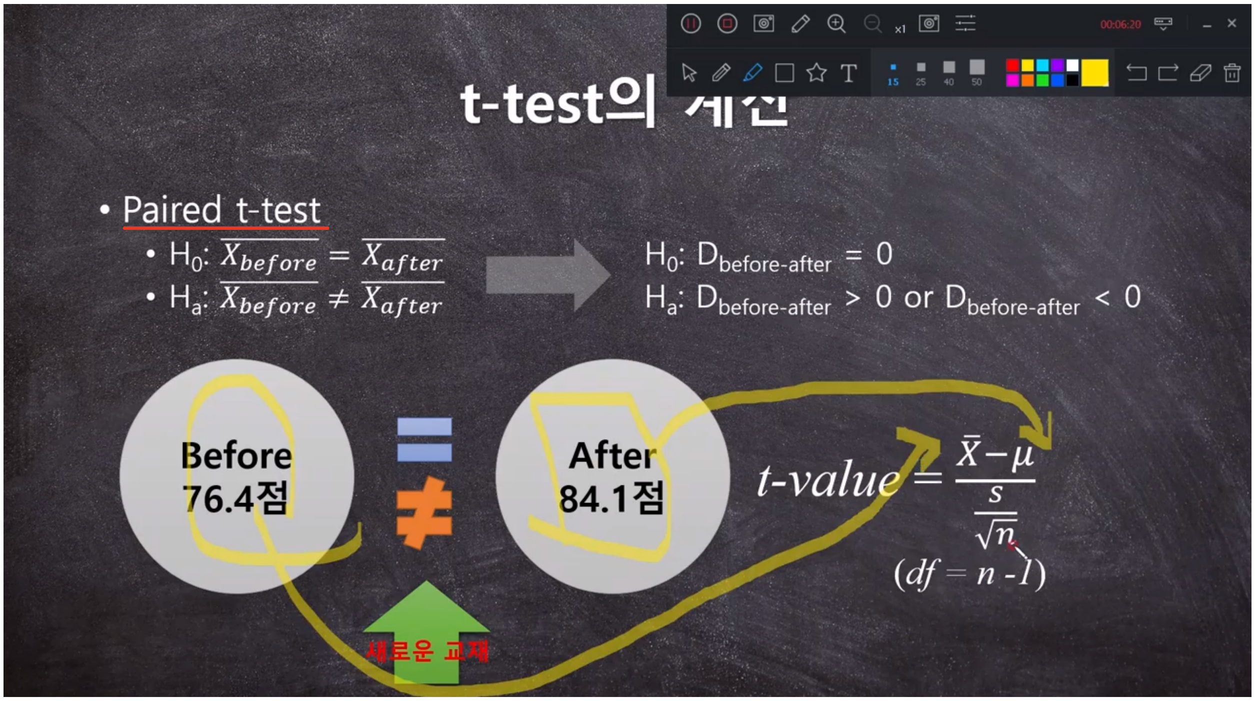
이렇게 같이 차이가 나더라도 그 정도를 파악하여 올바른 판단을 하게끔 도와주는 것이

바로 **T-test**입니다.

※ t-value 계산법.


t-test의 계산
• Two-sample t-test
• H_0: $\overline{X_a} = \overline{X_b}$
• H_a: $\overline{X_a} \neq \overline{X_b}$
H_0: $D_{a-b} = 0$
H_a: $D_{a-b} > 0$ or $D_{a-b} < 0$
A대학
178.5cm
B대학
179.9cm
$t\text{-}value = \frac{\overline{X_a} - \overline{X_b}}{\sqrt{\frac{s^2}{n_a} + \frac{s^2}{n_b}}}$
$s^2 = \frac{\sum(x_a - \overline{X_a})^2 + \sum(x_b - \overline{X_b})^2}{n_a + n_b - 2}$
$(df = n_a + n_b - 2)$


t-test의 계산
• One-sample t-test
• H_0: $\bar{X} = \mu$
• H_a: $\bar{X} \neq \mu$
H_0: $\bar{X} - \mu = 0$
H_a: $\bar{X} - \mu > 0$ or $\bar{X} - \mu < 0$
X대학
178.5cm
180.0cm
$t\text{-}value = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$
$(df = n\text{-}1)$


t-test의
• Paired t-test
• H_0: $\overline{X_{before}} = \overline{X_{after}}$
• H_a: $\overline{X_{before}} \neq \overline{X_{after}}$
H_0: $D_{before\text{-}after} = 0$
H_a: $D_{before\text{-}after} > 0$ or $D_{before\text{-}after} < 0$
Before
76.4점
=
≠
After
84.1점
$t\text{-}value = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$
$(df = n - 1)$
새로운 교재
00:06:20

요약 정리

- 검정 통계량: 통계적 가설의 진위여부를 검정하기 위해 표본으로 부터 계산하는 통계량 (표본 통계량의 2차 가공물)
- T-value의 의미: 차이 / 불확실도
  ↑ 추정량으로써의 불확실도. (보통, '표본 평균의 표준편차'를 사용)
- T-value를 수식으로 나타내면

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

- "충분히 큰 t-value"는 t-분포 상에서 결정되고, t-분포의 대략적인 형태는 모집단에서 두 표본집단을 추출하여 계산하는 과정을 거쳐 확인해볼 수 있다.

t-value가 커지려면 분자인 '$\bar{x}_1 - \bar{x}_2$'가 커야하고,

분모인 '$s_{\bar{x}_1 - \bar{x}_2}$'가 작아야함.