Deep Dive :
from EDA to CDA

통계의 방법론들과 숨겨진 직관들

D:plus

0. 강의 목표.

1. EDA와 CDA를 연결합니다.
2. 통계적인 기법들을 깊게 파봅니다 I : 방법 뒤에 숨겨진 직관을 추적합시다.
3. 통계적인 기법들을 깊게 파봅니다 II : 수학적인 구조도 뜯어봅니다.
4. 수식과 통계적인 표현과 현실의 언어를 연결합시다.

D:plus

목차

이 강의는 The First Data Analysis 이후 수강하기를 권장합니다.

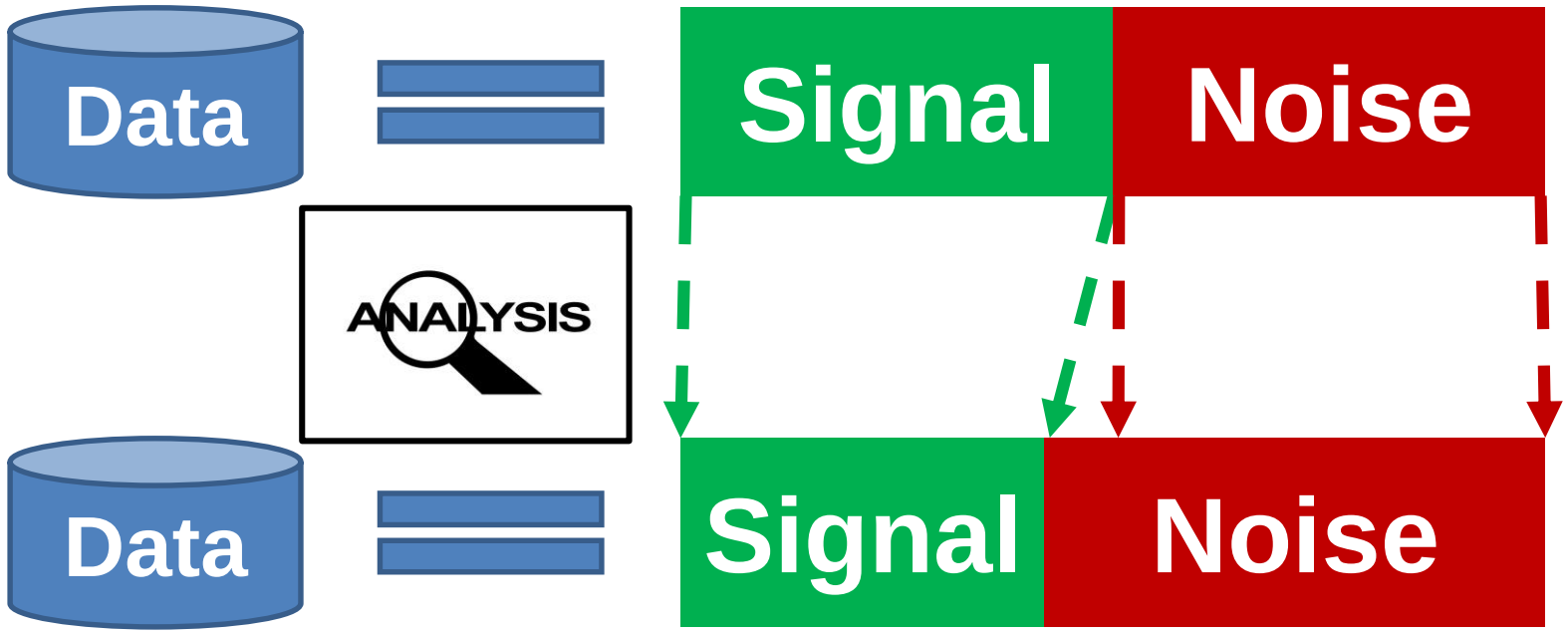
D:plus

1. Descriptive Type :
요약의 힘

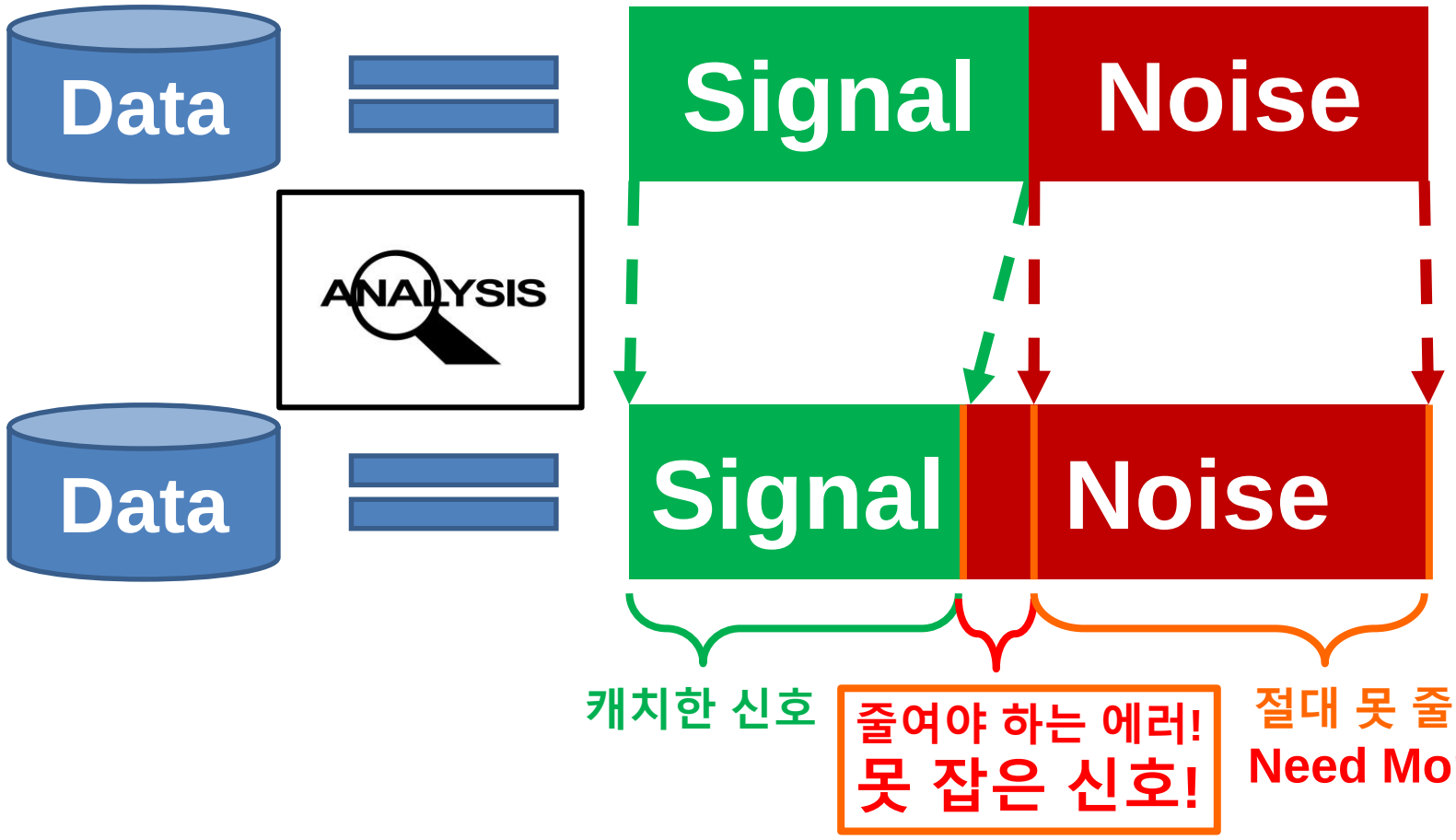
1. Descriptive Type : 요약의 힘

Data가 많아야 좋다? → 데이터 자체에 Noise가 상대적으로 작다. → 물론, Signal을 캐치하는 난이도는 높아진다.

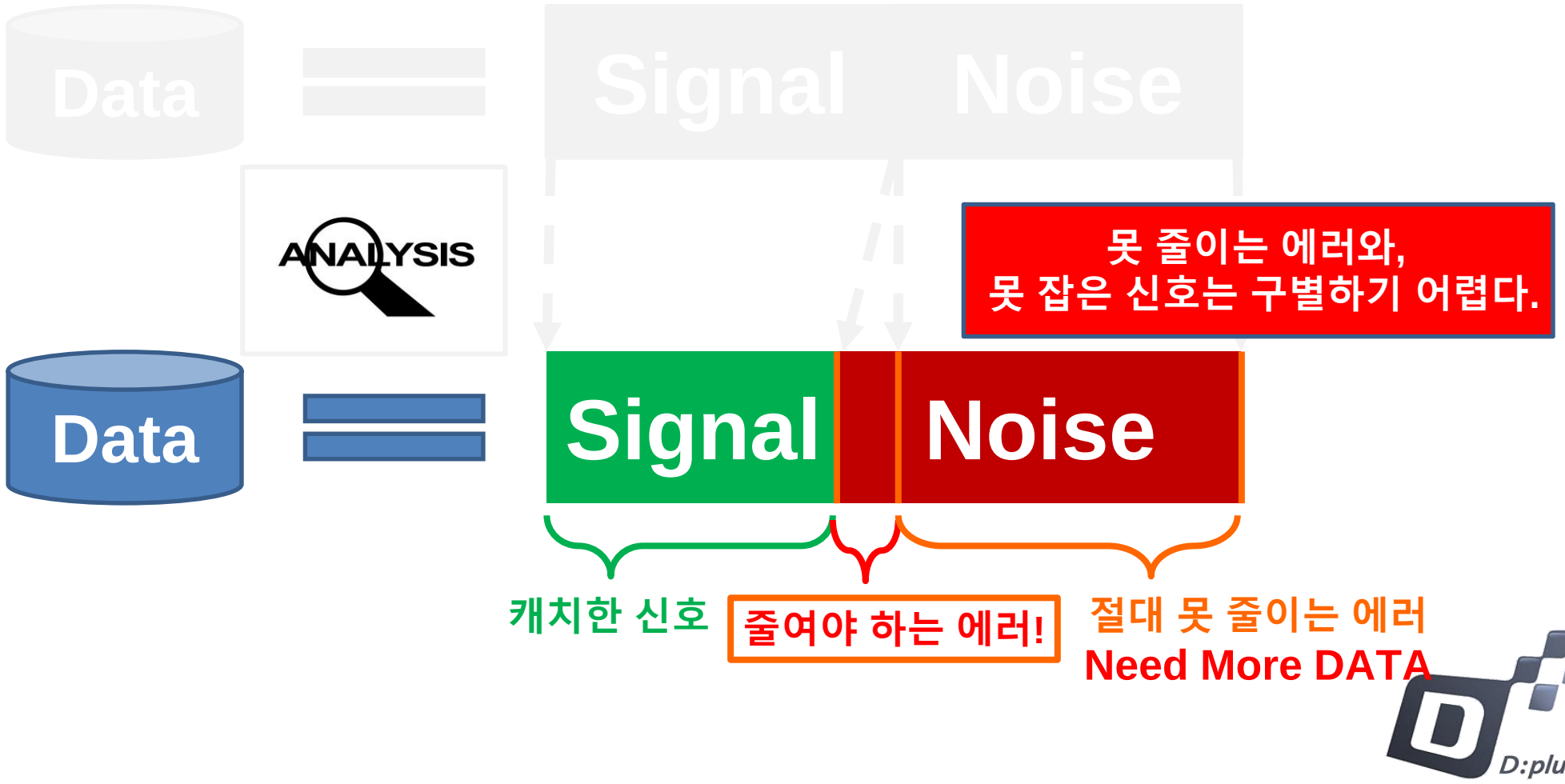
1. Descriptive Type : 요약의 힘



1. Descriptive Type : 요약의 힘



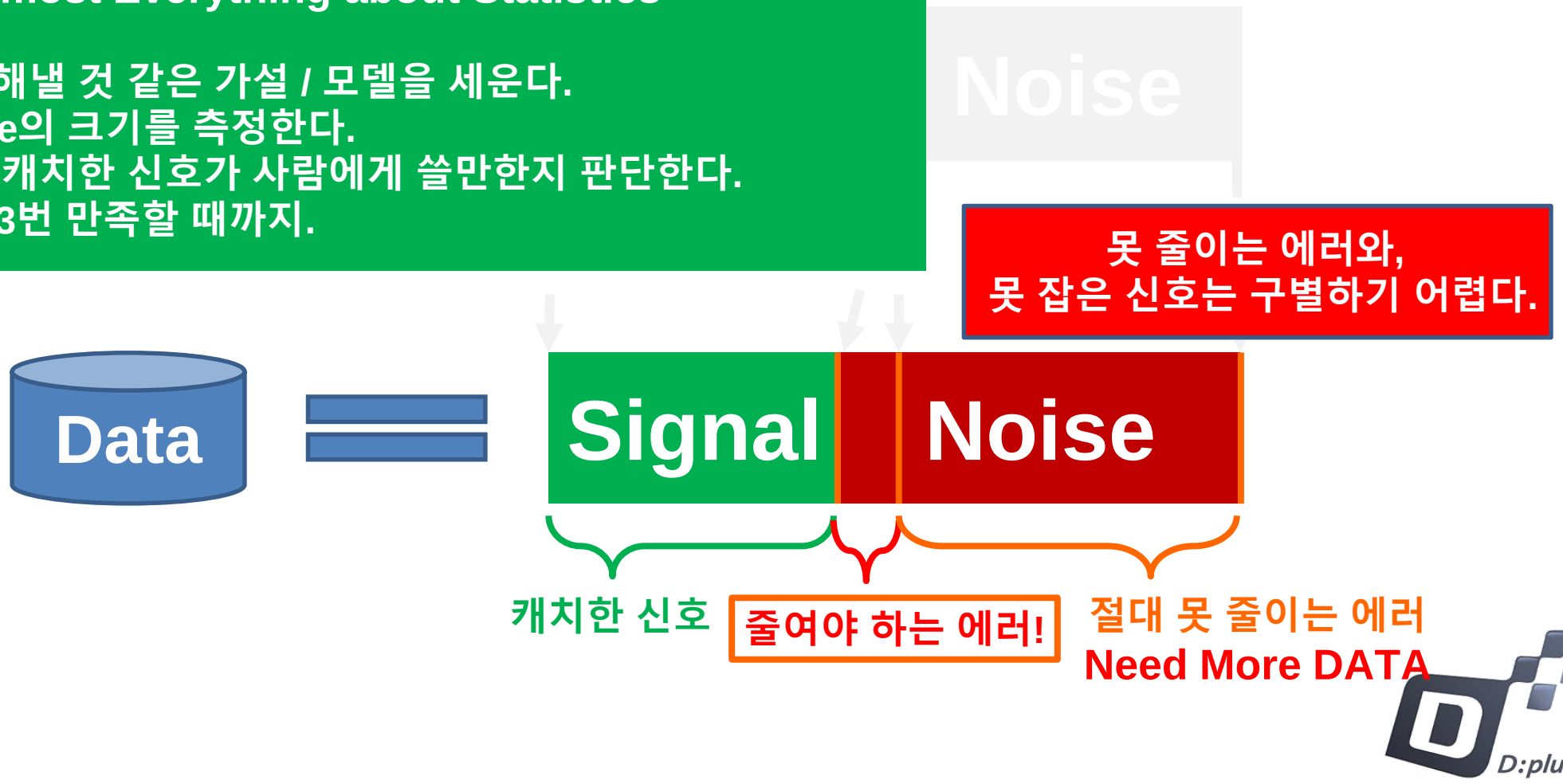
1. Descriptive Type : 요약의 힘



1. Descriptive Type : 요약의 힘

Almost Everything about Statistics

1. Signal을 캐치해낼 것 같은 가설 / 모델을 세운다.
2. Signal과 Noise의 크기를 측정한다.
3. 서로 비교하여 캐치한 신호가 사람에게 쓸만한지 판단한다.
4. 1로 돌아가라. 3번 만족할 때까지.



1. Descriptive Type : 요약의 힘

Almost Everything about Statistics

1. Signal을 캐치해낼 것 같은 가설 / 모델을 세운다.
2. **Signal과 Noise의 크기를 측정한다.**
3. 서로 비교하여 캐치한 신호가 사람에게 쓸만한지 판단한다.
4. 1로 돌아가라. 3번 만족할 때까지.

Noise

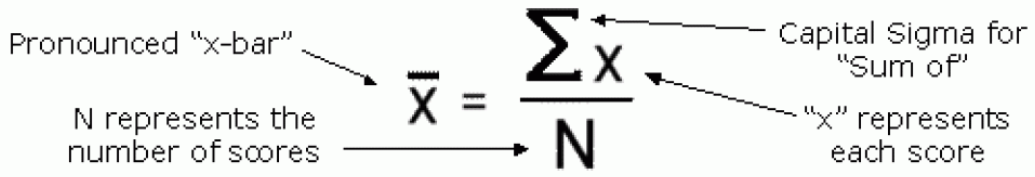
그래도 최대한 Signal을 잡아야 하는 것이 우리의 운명.

D:plus

1. Descriptive Type : 요약의 힘

Q. 평균을 보라고들 하는데, 왜 평균을? 어디에 쓸 수 있지?

1. **Mean** = Sum of scores divided by the number of scores (often referred to as the statistical average)



$$\bar{x} = \frac{\sum x}{N}$$

2. **Median** = Middle Most Number

$$M_d$$

3. **Mode** = Most Frequently Occurring Number

$$M_o$$

1. Descriptive Type : 요약의 힘

Q. 평균을 보라고들 하는데, 평균의 한계는?

10명의 학생들 중
9명의 학생의 용돈 최대값은 50000원 이다.

1명은 용돈이 1,000,000원이다.

만원, 만원, 만원, 이만원, 삼만원, 사만원, 오만원, 삼만원, 이만원, 백만원.

D:plus

1. Descriptive Type : 요약의 힘

Q. 평균을 보라고들 하는데, 다른 방법은 없나?

Compute a 10% trimmed mean:

15, 17, 18, 20, 20, 25, 30, 32, 36, 60

- Delete the top and bottom 10%
- New data list:

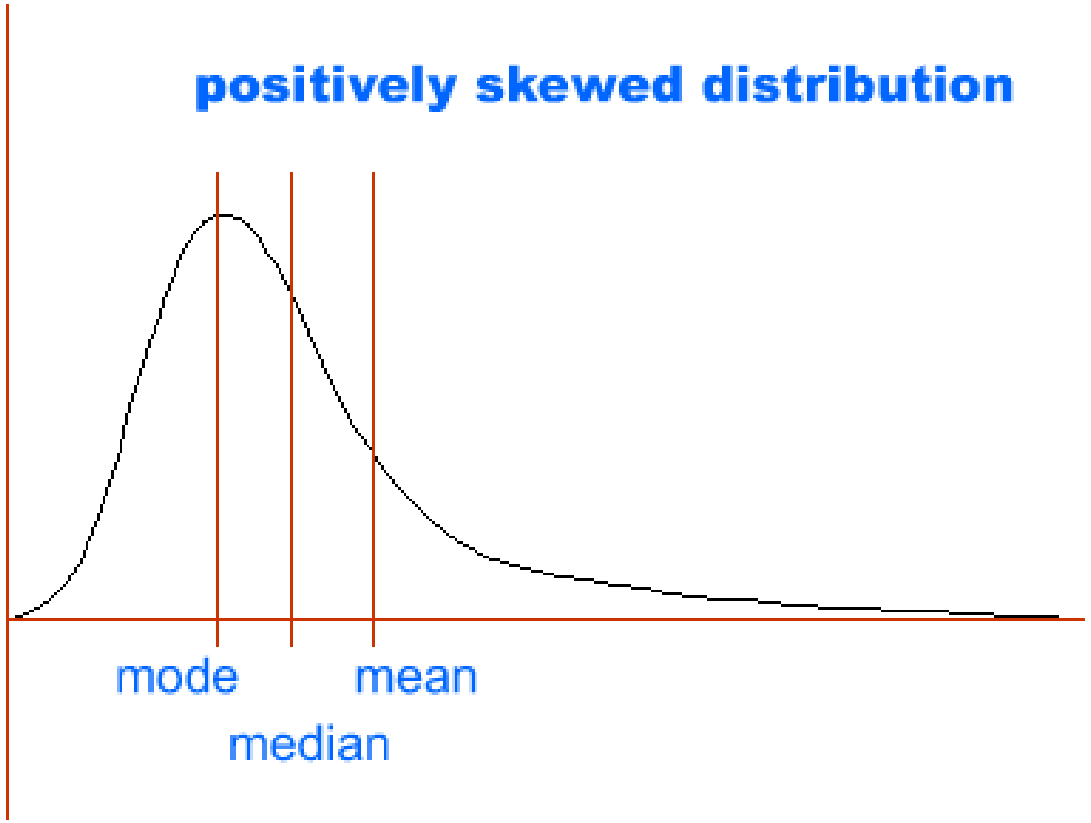
  17, 18, 20, 20, 25, 30, 32, 36
- 10% trimmed mean =

$$\frac{\sum x}{n} = \frac{198}{8} \approx 24.8$$

D:plus

1. Descriptive Type : 요약의 힘

Q. 최빈값(Mode)라는 것도 있다. 장점은? 그리고 한계는?



1. Descriptive Type : 요약의 힘

Q. 평균을 보라고들 하는데, 평균의 한계는?

	1월	2월	3월	4월	5월	6월	7월
철수	-3	-4	5	6	-2	6	-1
영희	-1	2	-1	3	-2	3	3

최고의 주식 투자 전문가 철수와 영희.

당신은 누구를 선택하시겠습니까.

철수와영희.xlsx 파일을 열어서 관찰해봐도 좋다.

D:plus

1. Descriptive Type : 요약의 힘

Q. 평균만 볼 거야? 최빈값? 중앙값?

	1월	2월	3월	4월	5월	6월	7월
철수	-3	-4	5	6	-2	6	-1
영희	-1	2	-1	3	-2	3	3

관찰한 것들을 다 말해보자.

D:plus

1. Descriptive Type : 요약의 힘

편차 : 평균으로 부터 얼마나 차이가 나는가?

편차!	1월	2월	3월	4월	5월	6월	7월
철수	-4	-5	4	5	-3	5	-2
영희	-2	1	-2	2	-3	2	2

이것이 도대체 어떻게 분산으로, 표준편차로 발전 한 것인가!?

D:plus

1. Descriptive Type : 요약의 힘

편차 : 평균으로 부터 얼마나 차이가 나는가?

편차!	1월	2
철수	-4	
영희	-2	

$$\text{variance} = \sigma^2 = \frac{\sum (x_r - \mu)^2}{n}$$

$$\text{standard deviation} \quad \sigma = \sqrt{\frac{\sum (x_r - \mu)^2}{n}}$$

$\mu = \text{mean}$

	7월
5	-2
2	2

1. Descriptive Type : 요약의 힘

중간 요약 (기본 관점)

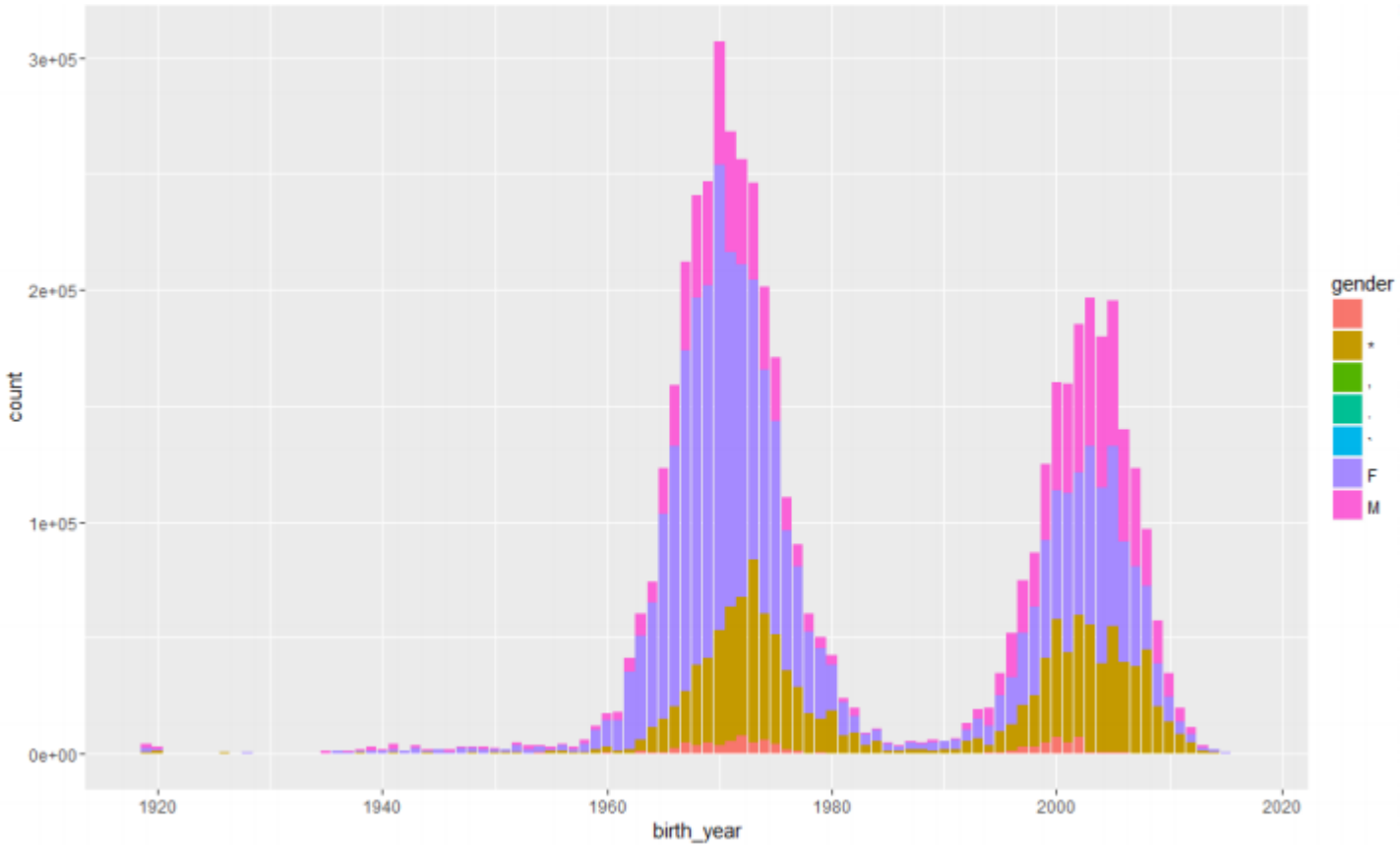
Central Tendency : 신호를 요약하는 수단들.
Variability : 노이즈를 요약하는 수단들.

요약을 하나의 숫자로만 해서는 안 되는 이유:
ㄱ. 요약 수단에 따라 하나의 관점만 부각됨.
ㄴ. 내가 원하는 신호가 의미가 있으려면 오차(노이즈)도 같이 확인해야 함.

이제, 사람들은 숫자로만 요약하는 것을 넘어서기 시작한다.

1. Descriptive Type : 요약의 힘

대표값으로만 관찰하면 보이지 않는 것들. 눈으로 볼 필요가 있다.



출생 년도 대비 도서관 이용자 수 카운트.
2016년 07월 분석

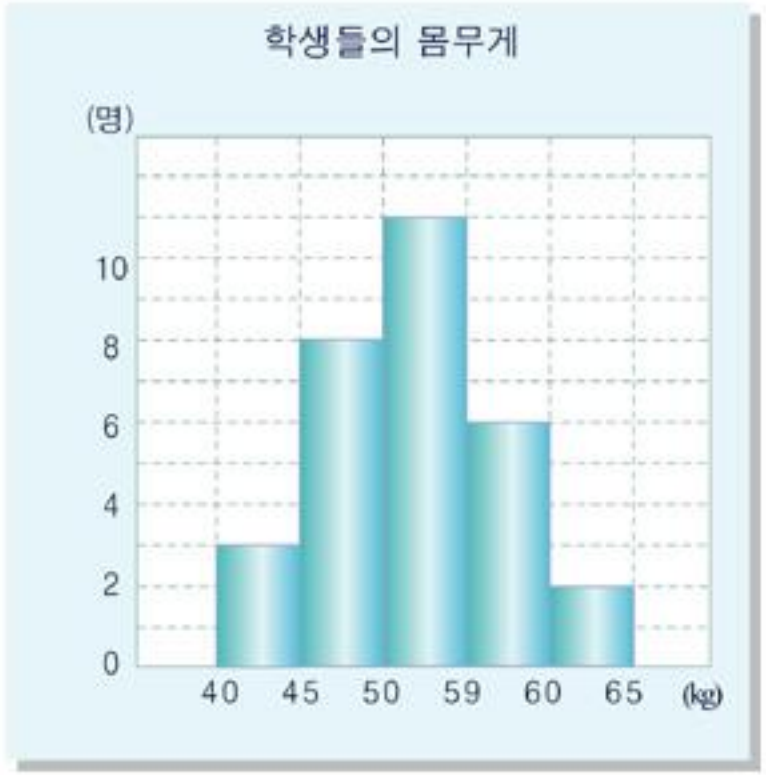
Q. 왜 이봉으로 나뉜 걸까?!

Q. 보통 ML에선 클러스터링을 권장한다. 정말 저거 잘라서 나눠도 괜찮을까?

D:plus

1. Descriptive Type : 요약의 힘

Histogram. 혹은 Bar plot



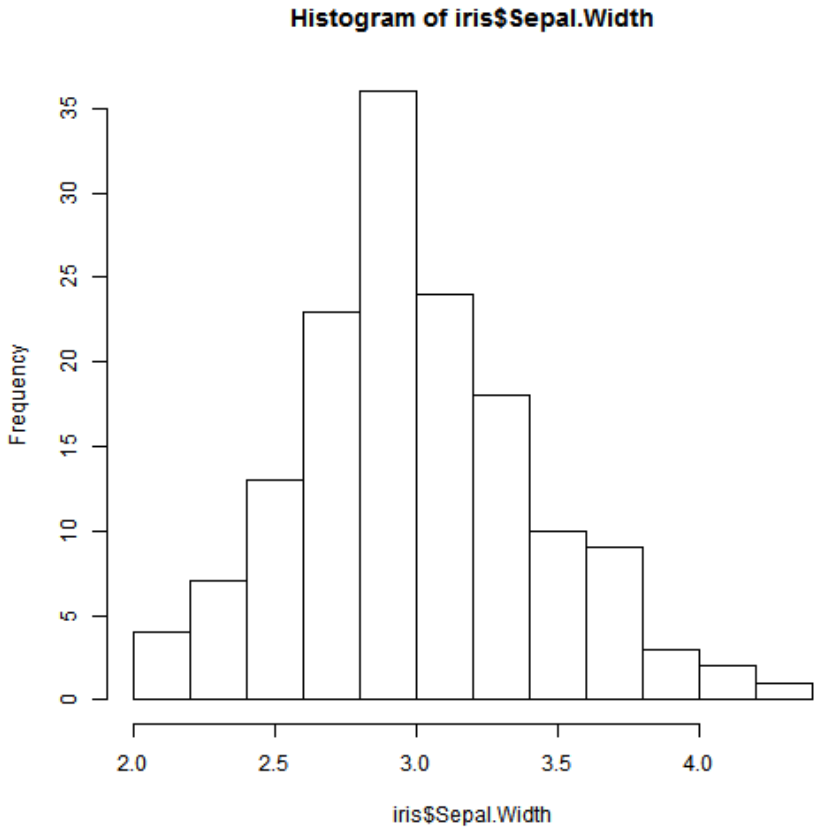
총 30명의 학생 몸무게를 가지고 만든 히스토그램.

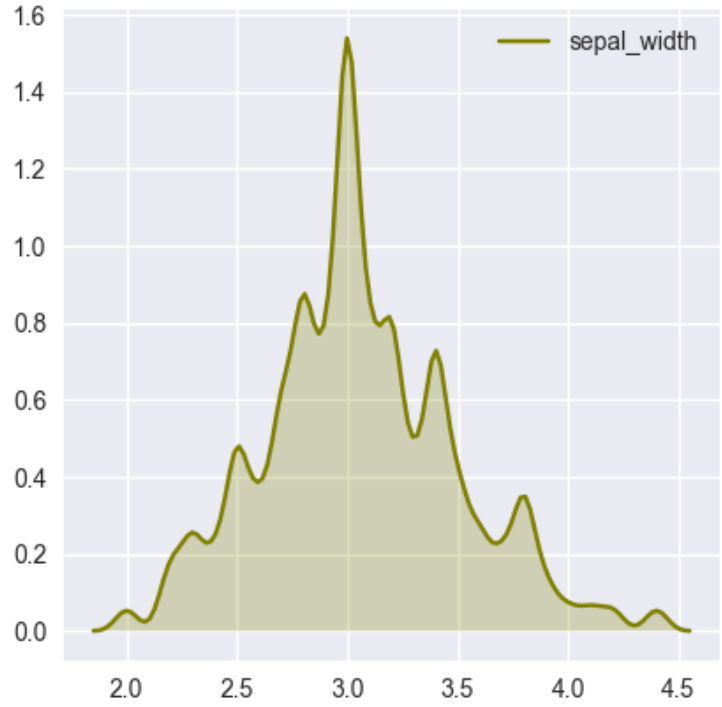
임의의 학생 한 명이 40kg~45kg 사이일 확률은?

1. Descriptive Type : 요약의 힘

Histogram. 혹은 Bar plot

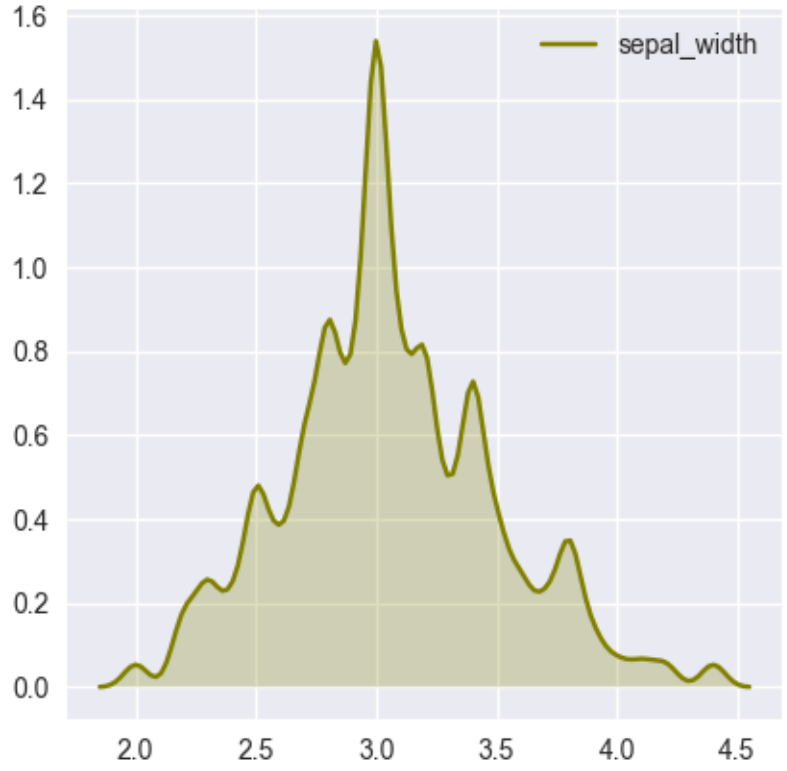
확률변수? 확률밀도? 분포? Density plot?





1. Descriptive Type : 요약의 힘

변수 하나를 관찰하는 끝판 왕. Density plot



Trade off !

설명이 간결하다 vs 정보가 풍부하다

D:plus

1. Descriptive Type : 요약의 힘

Summary.

1. 눈으로 보면 강력하다!
2. 요약이 강해지면 설명이 쉬워지지만, 정보손실(노이즈)도 커진다.
3. 요약을 덜 할수록, 신호는 많이 잡히지만 설명이 쉽지 않다.

--Remind EDA

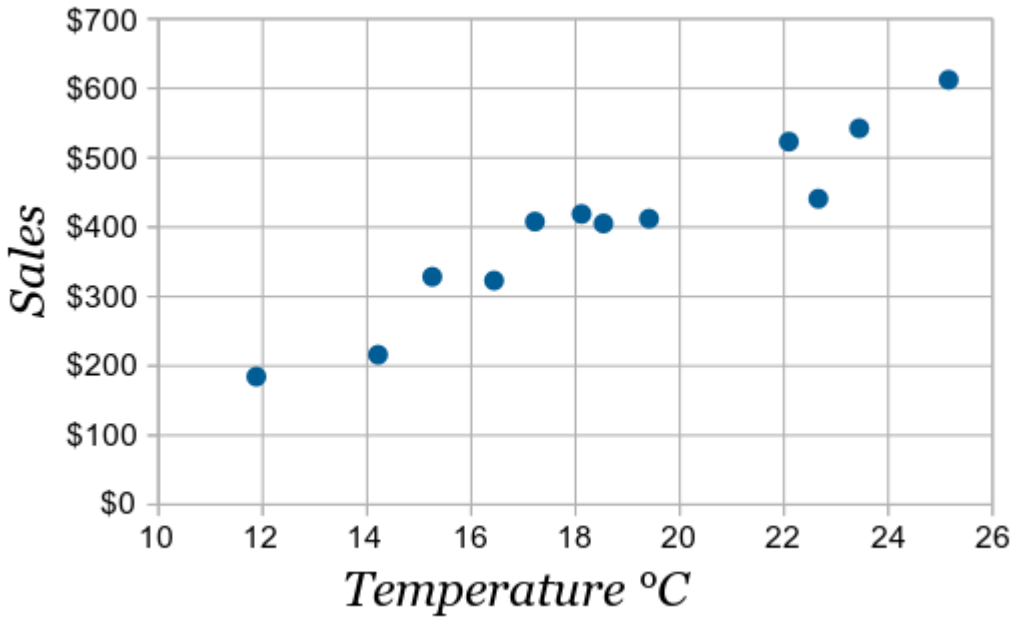
방법론에 흔들리면 안 된다.
요약된 결과를 보고 의문을 품고 가설을 세울 것.
요약된 결과를 보고 부족한 부분을 찾을 것.

D:plus

2. Associative Type I :
Correlation 뜯어보기

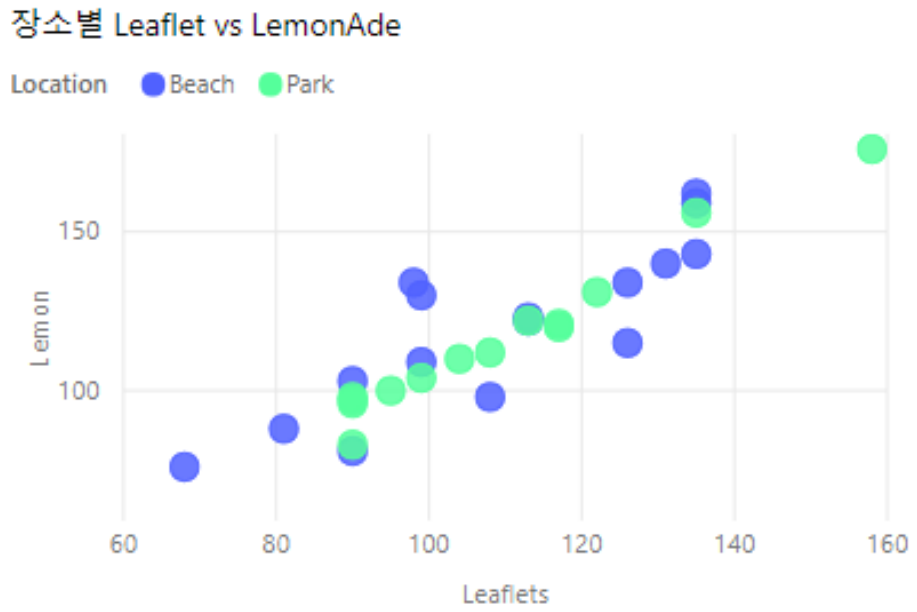
2.3 차근차근II : Associative type

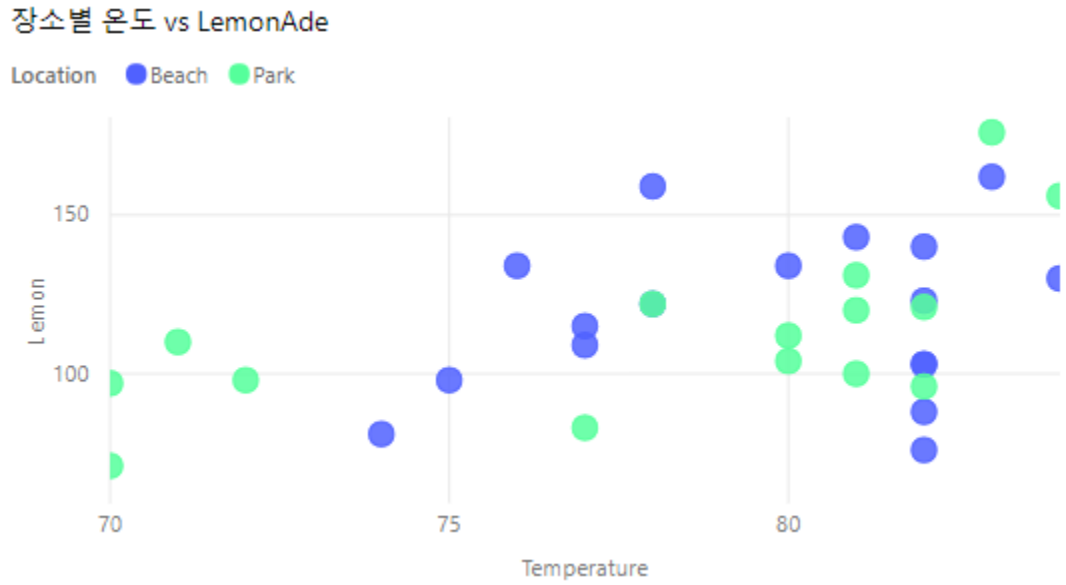
상관 분석 (Correlation Analysis)에 대해서 이야기 해보자.



2. Associative Type I : Correlation 뜯어보기

연속형 변수로 연속형 변수를 설명하려는 시도 : Scatter Plot

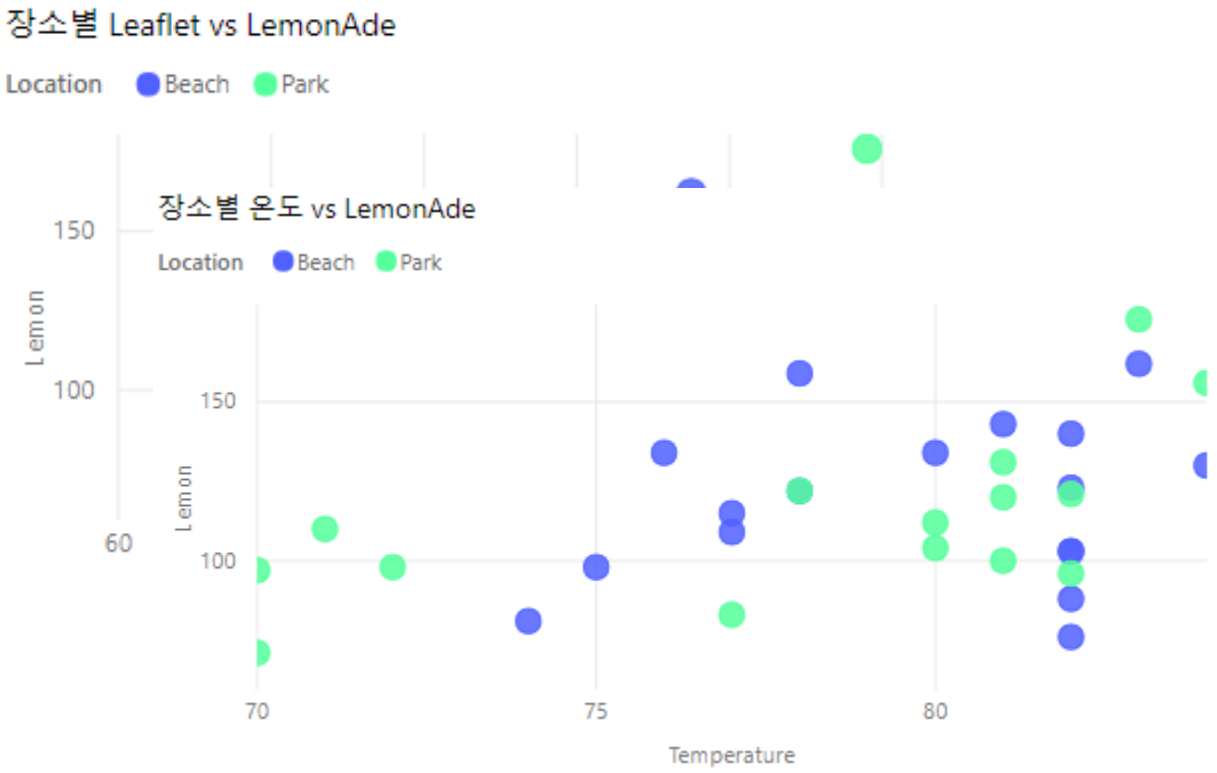




D:plus

2. Associative Type I : Correlation 뜯어보기

생각해볼 법한 시도들



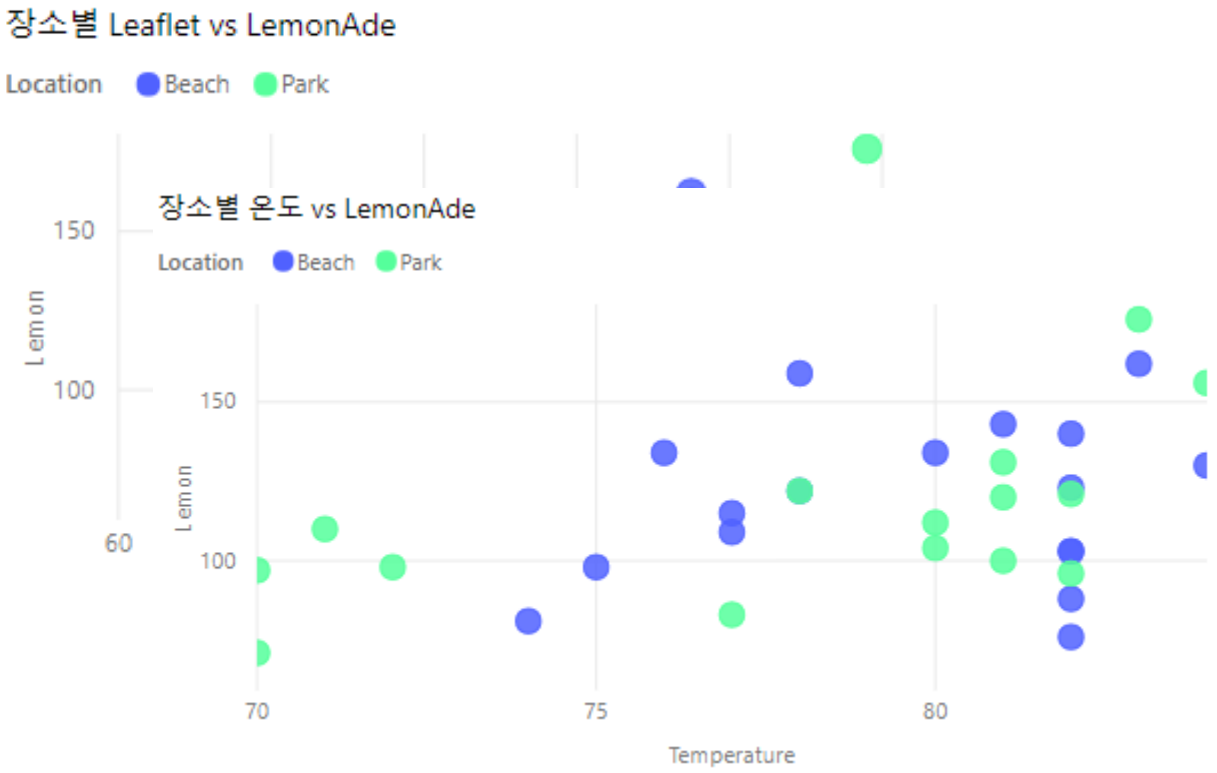
무슨 시도를 해보겠습니까?

1. 나는 나의 길을 간다. 자를 들어 가장 그럴싸한 선을 그리고 기울기를 구하겠다.
2. 너만 믿는다 Linear Regression.
3. 일단 눈으로 관계를 확인해본다.
   선형성이 있는지, 다른 관계가 있는지.

D:plus

2. Associative Type I : Correlation 뜯어보기

생각해볼 법한 시도들



무슨 시도를 하

분석가의 출발선에 서게 된 그대에게 건배를.

1. 나는 나의 길을 간다. 자를 들어 가장 그럴싸한 선을 그리고 기울기를 구하겠다.
2. 너만 믿는다 Linear Regression.

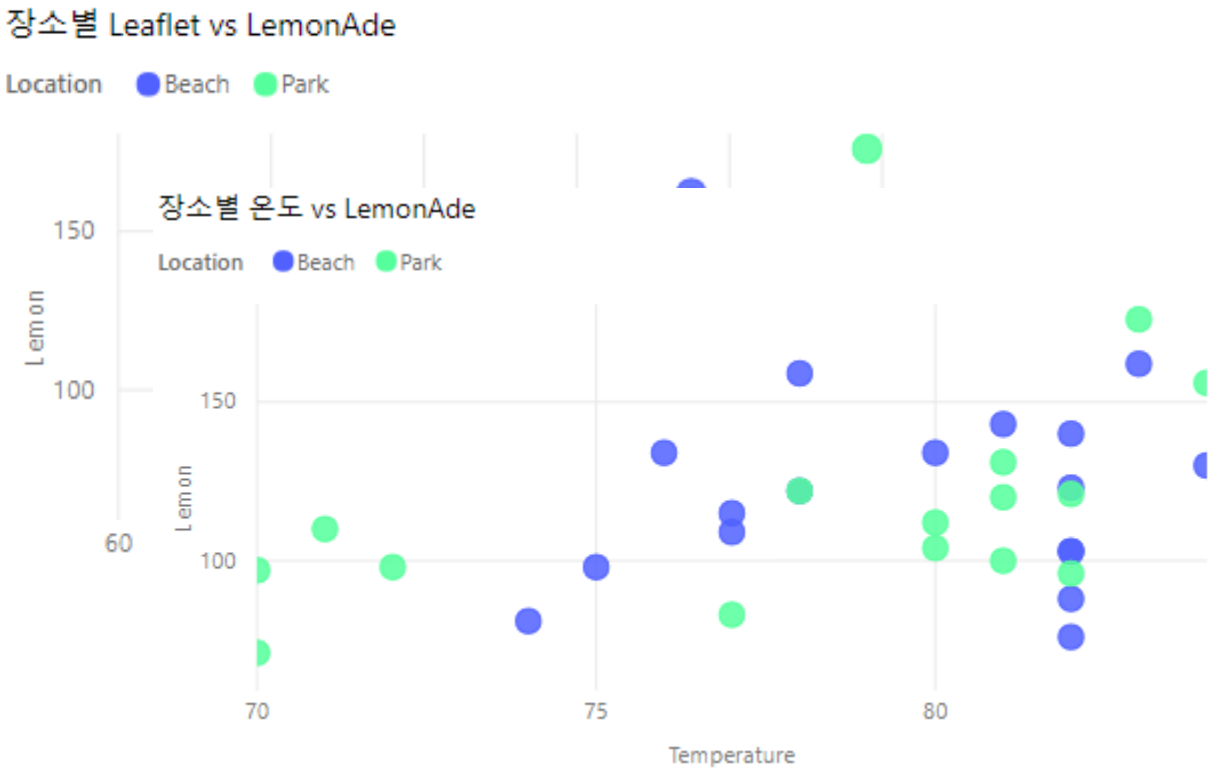
논리적인 순서로는 나중 단계. 하지만...?

3. 일단 눈으로 관계를 확인해본다.
   선형성이 있는지, 다른 관계가 있는지.

D:plus

2. Associative Type I : Correlation 뜯어보기

생각해볼 법한 시도들



무슨 시도를 해보겠습니까?

1. 나는 나의 길을 간다. 자를 들어 가장 그럴싸한 선을 그리고 기울기를 구하겠다.

2. 너만 믿는다 Linear Regression.

3. 일단 눈으로 관계를 확인해본다.
선형성이 있는지, 다른 관계가 있는지.

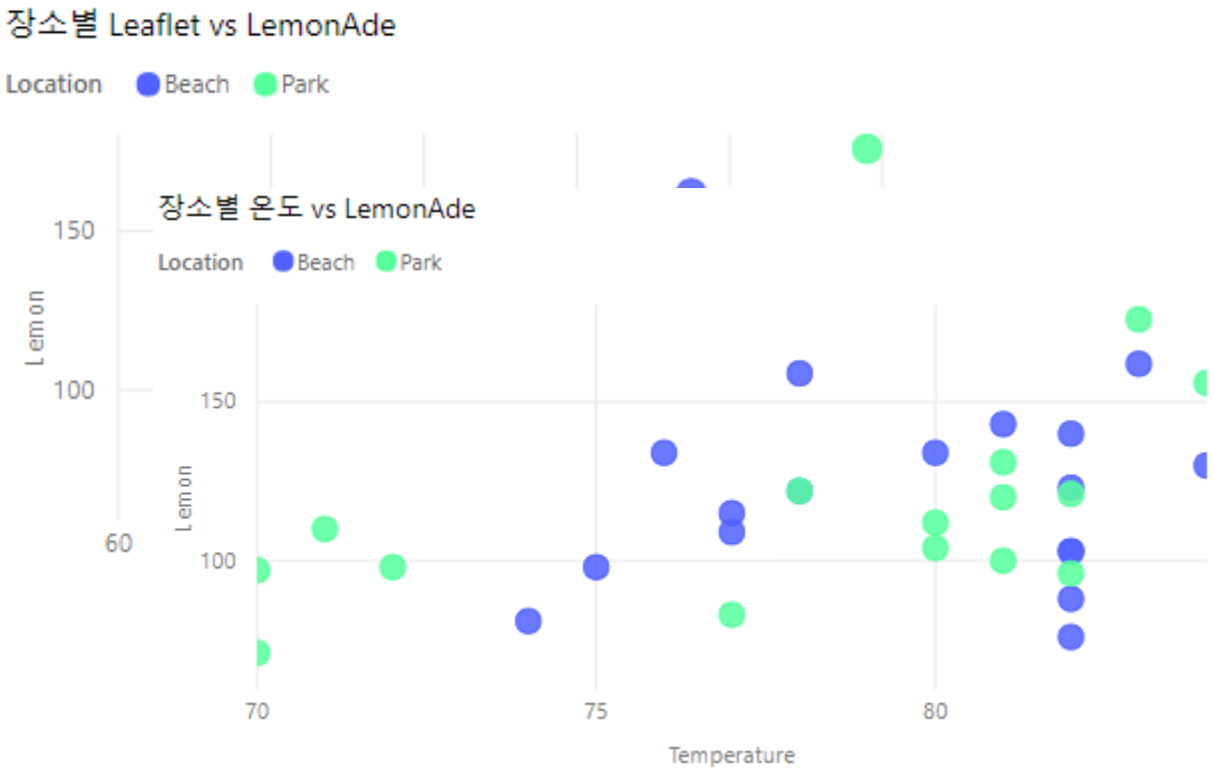
가장 설명하기 쉬운 관계는 Linearity!

Ex> 유동인구가 늘어날 수록, 매출도 비례해 오르더라.

D:plus

2. Associative Type I : Correlation 뜯어보기

생각해볼 법한 시도들



가장 설명하기 쉬운 관계는 Linearity!

Ex> 유동인구가 늘어날 수록, 매출도 비례해 오르더라.

두 변수 사이의 Linearity를 판단하기에 가장 간단한 도구

Pearson Correlation Coefficient!

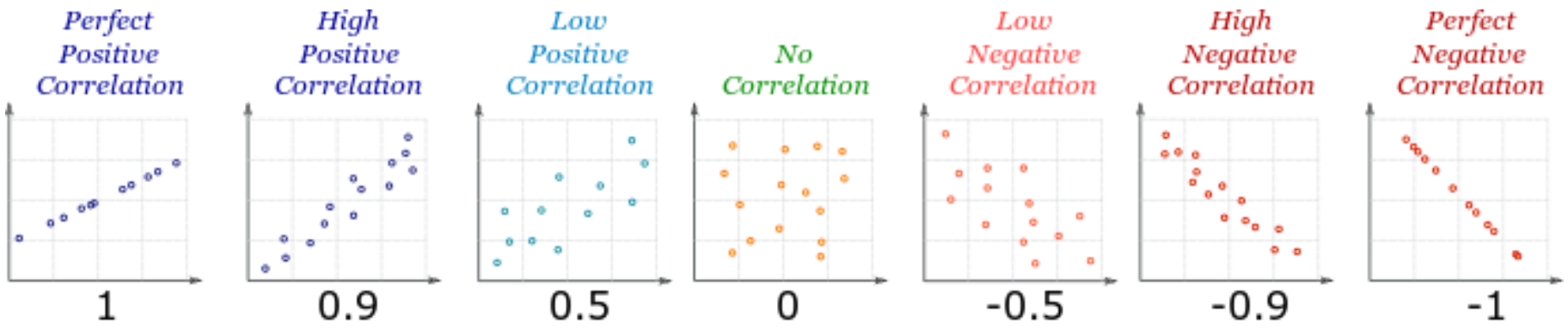
D:plus

2. Associative Type I : Correlation 뜯어보기

Pearson Correlation Coefficient!

a.k.a

상관계수

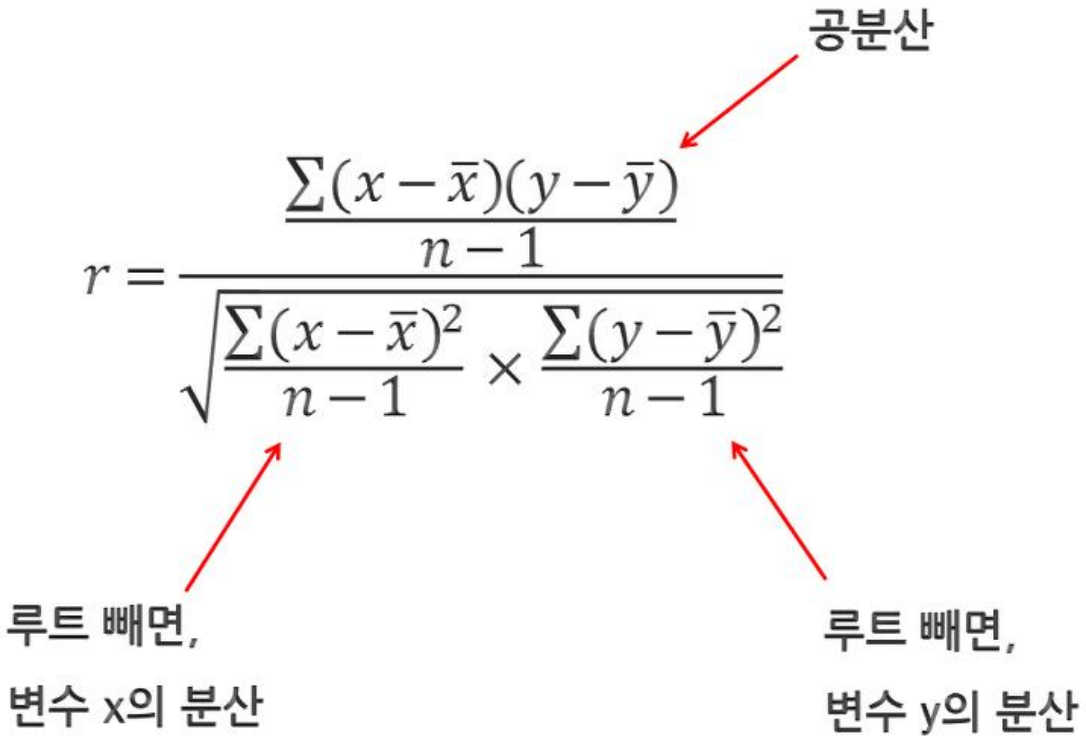


선형적인 관계가 얼마나 강한지 미리 알 수 있다!

D:plus

2. Associative Type I : Correlation 뜯어보기

상관 계수 : 조금만 더 들어가보자.



분자를 뜯어서 이해해보자.

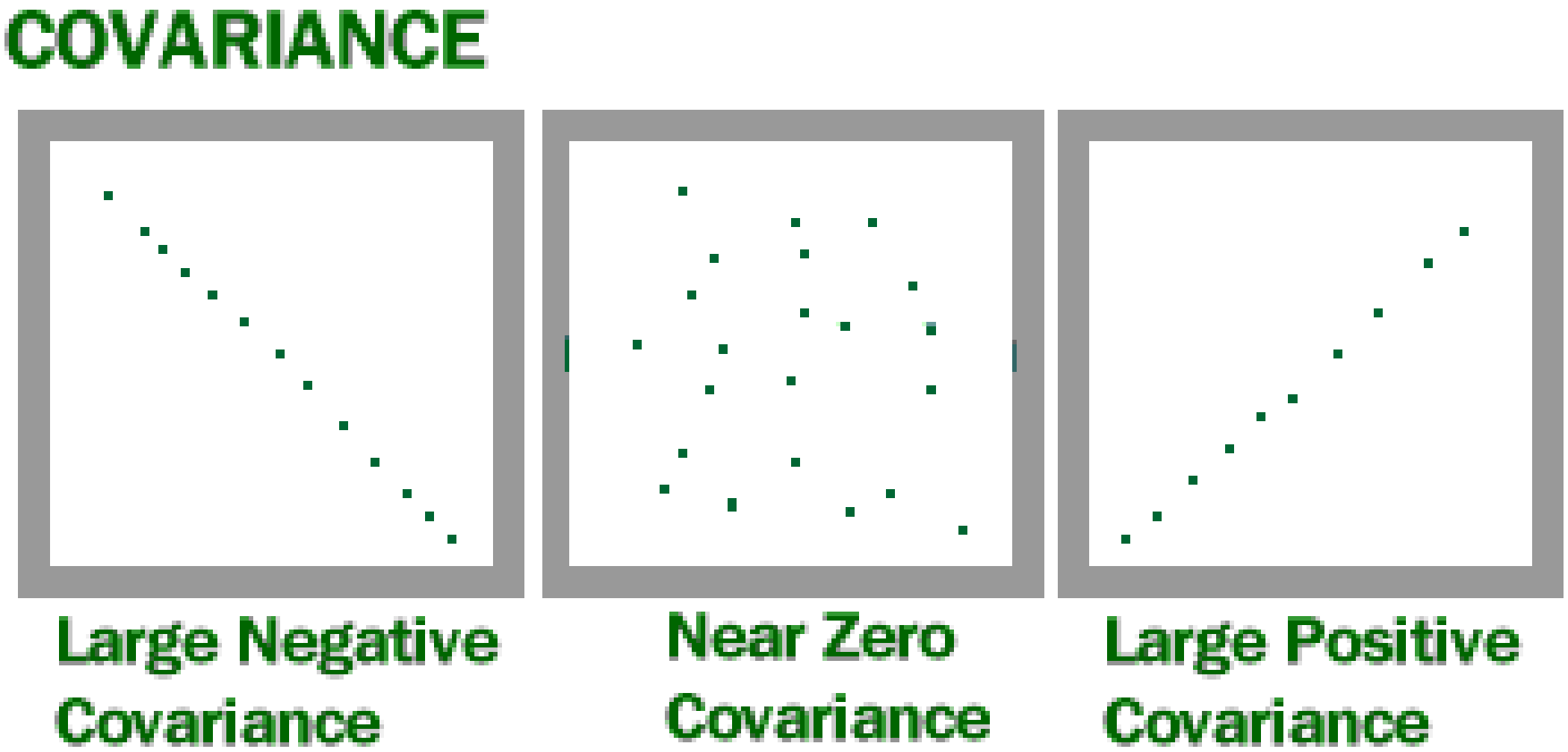
1. **공분산이 언제 0에 가까운가?**
2. **공분산이 언제 +로 커지는가?**
3. **공분산이 언제 –로 커지는가?**

흔히들 이야기 하는 최대값이 1에 대한 증명은
http://freshrimpsushi.tistory.com/57 링크 참고.

하지만 우리는, Intution을 더 찾아보자.

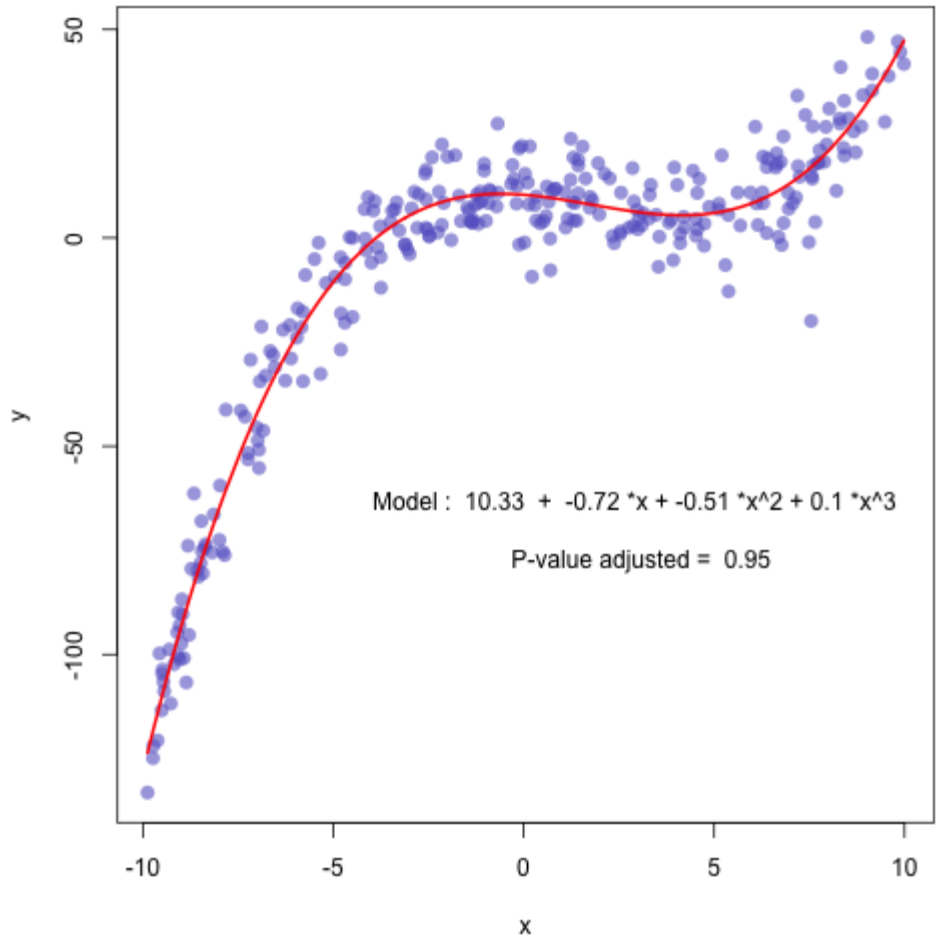
D:plus

2. Associative Type I : Correlation 뜯어보기



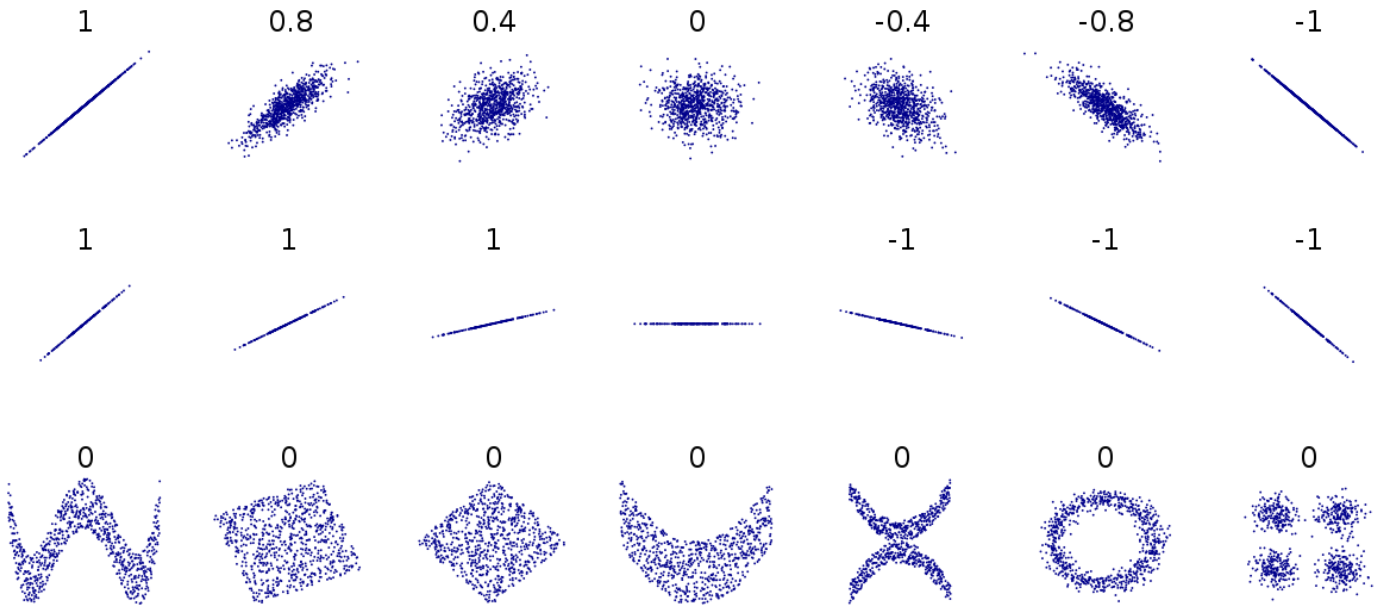
2. Associative Type I : Correlation 뜯어보기

상관 계수 : 한계점



2. Associative Type I : Correlation 뜯어보기

한계점



3. Associative Type II :

Chi-Squared test 뜯어보기

3. Associative Type II : Chi-Squared test 뜯어보기

명목형 변수들 사이에 관계가 있지 않는가? : 분할표 (Contingency Table)

	관측된 값들		
	당뇨	정상	전체
비만체중	10(40%)	10(13.3%)	20(20%)
정상체중	15(60%)	65(86.7%)	80(80%)
전체	25(100%)	75(100%)	100

떠오르는 생각들을 전부 적어보자.

D:plus

3. Associative Type II : Chi-Squared test 뜯어보기

변수들을 같이 본다면. 비연속 and 비연속

	관측된 값들			확률적으로 독립일때 기대되는 값	
	당뇨	정상	전체	당뇨	정상
비만체중	10(40%)	10(13.3%)	20(20%)		
정상체중	15(60%)	65(86.7%)	80(80%)		
전체	25(100%)	75(100%)	100	25	75

D:plus

3. Associative Type II : Chi-Squared test 뜯어보기

변수들을 같이 본다면. 비연속 and 비연속

	관측된 값들			확률적으로 독립일때 기대되는 값	
	당뇨	정상	전체	당뇨	정상
비만체중	10(40%)	10(13.3%)	20(20%)	5	15
정상체중	15(60%)	65(86.7%)	80(80%)	20	60
전체	25(100%)	75(100%)	100	25	75

D:plus

3. Associative Type II : Chi-Squared test 뜯어보기

카이제곱 검정

	관측된 값들			확률적으로 독립일때 기대되는 값	
	당뇨	정상	전체	당뇨	정상
비만 체중	10	10	20	5	15
정상 체중	15	65	80	20	60
전체	25	75	100	25	75

$$\chi^2 = \sum \frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}} = \frac{(+5)^2}{5} + \frac{(-5)^2}{20} + \frac{(-5)^2}{15} + \frac{(+5)^2}{60} = 8.33$$

분자의 의미에 대해서 고찰해보자.

D:plus

3. Associative Type II : Chi-Squared test 뜯어보기

카이제곱 검정

	관측된 값들			확률적으로 독립일때 기대되는 값	
	당뇨	정상	전체	당뇨	정상
비만 체중	10	10	20	5	15
정상 체중	15	65	80	20	60
전체	25	75	100	25	75

$$\chi^2 = \sum \frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}} = \frac{(+5)^2}{5} + \frac{(-5)^2}{20} + \frac{(-5)^2}{15} + \frac{(+5)^2}{60} = 8.33$$

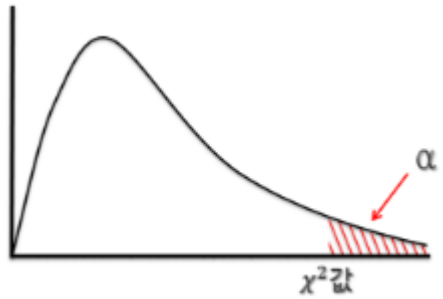
지금은 일단, 참고 넘어가는 것.

자유도 계산
df = (nrows-1)(ncols-1)

3. Associative Type II : Chi-Squared test 뜯어보기

카이제곱 분포표

ν \ α	0.995	0.99	0.975	0.95	0.9	0.5	0.1	0.05	0.025	0.01	0.005
1	0.00004	0.0002	0.001	0.004	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.43	104.21
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.15	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17



D:plus

4. Comparative type :

T-test부터 ANOVA살짝

4. Comparative Type : T-test부터 ANOVA살짝

비교는 강력하다.

4. Comparative Type : T-test부터 ANOVA살짝

비교. 비교. 비교.

Control

Variation

Conversion Rate

5.8% -> 10.2%

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

그 방법으로는, 가설검정을 사용한다.

대부분의 귀무가설 H0 : 보수적인 입장. 차이가 없다. 변화가 없다. 등등

대부분의 대립가설 H1 : 우리가 바라는 무언가. 차이가 있다. 변화가 있다. 등등.

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

비교할 일은 매우 많다! 그 방법으로는, 가설검정을 사용한다.

대부분의 귀무가설 H0 : 보수적인 입장. 차이가 없다. 변화가 없다. 등등

대부분의 대립가설 H1 : 우리가 바라는 무언가. 차이가 있다. 변화가 있다. 등등.

핵심 IDEA

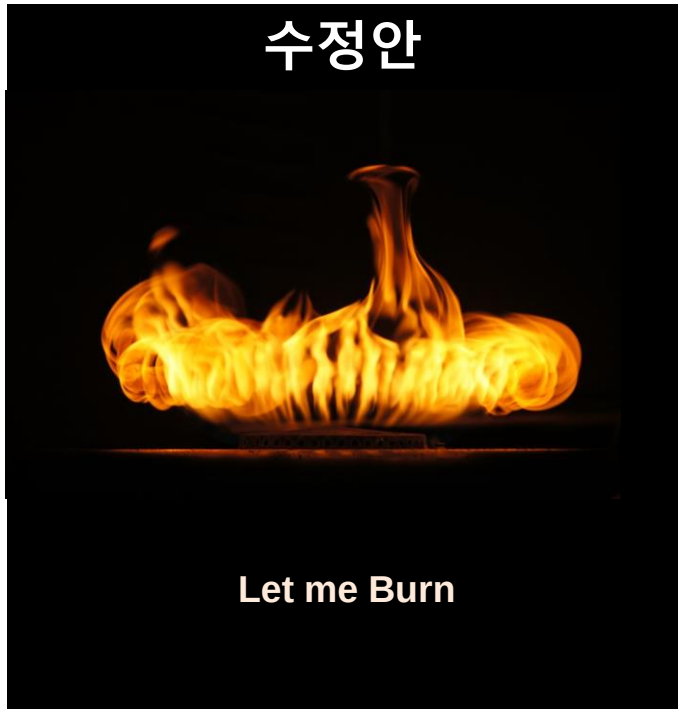
귀무가설이 참이라고 했을 때,
이런 데이터가 관찰이 될 확률은?

Ex> 주사위가 공평하다는데, 20번 중 10번이 6이 나올 확률은? 3번 6이 나올 확률은?

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

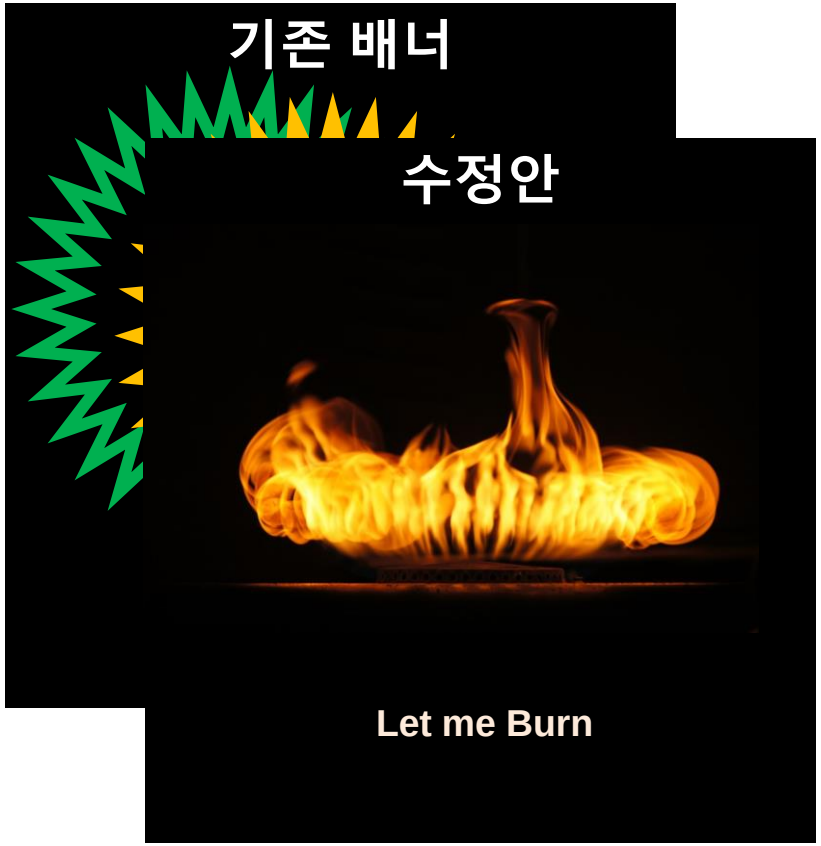
간단한 예제부터!



D:plus

4. Comparative Type : T-test부터 ANOVA살짝

T-test!



귀무 가설, 기존 입장, 보수적인 입장, 현재의 기준

H_0 : 기존 배너든 바꾸든 일 평균 방문수에 차이 없음

H_1 : $\mu_{new} = \mu_{old}$ // $\mu_{new} - \mu_{old} = 0$

대립가설, 새로운 입장

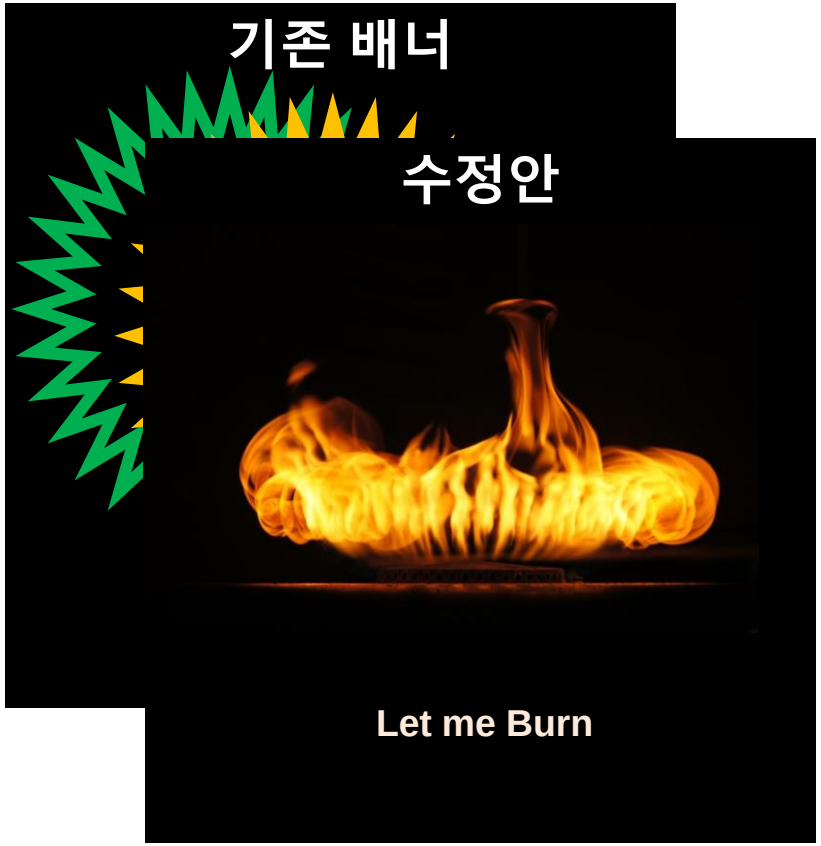
H_0 : 수정안의 일 평균 방문수가 더 나을 거야

H_1 : $\mu_{new} > \mu_{old}$ // $\mu_{new} - \mu_{old} > 0$

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

T-test!



귀무 가설, 기존 입장, 보수적인 입장, 현재의 기준

H_0 : 기존 배너든 바꾸든 일 평균 방문수에 차이 없음

H_1 : $\mu_{new} = \mu_{old}$ // $\mu_{new} - \mu_{old} = 0$

대립가설, 새로운 입장

H_0 : 수정안의 일 평균 방문수가 더 나을 거야

H_1 : $\mu_{new} > \mu_{old}$ // $\mu_{new} - \mu_{old} > 0$

$$t - statistics = \frac{[\mu_{new} - \mu_{old}]_{from\ Data} - [\mu_{new} - \mu_{old}]_{from\ H_0}}{SE(new, old)}$$

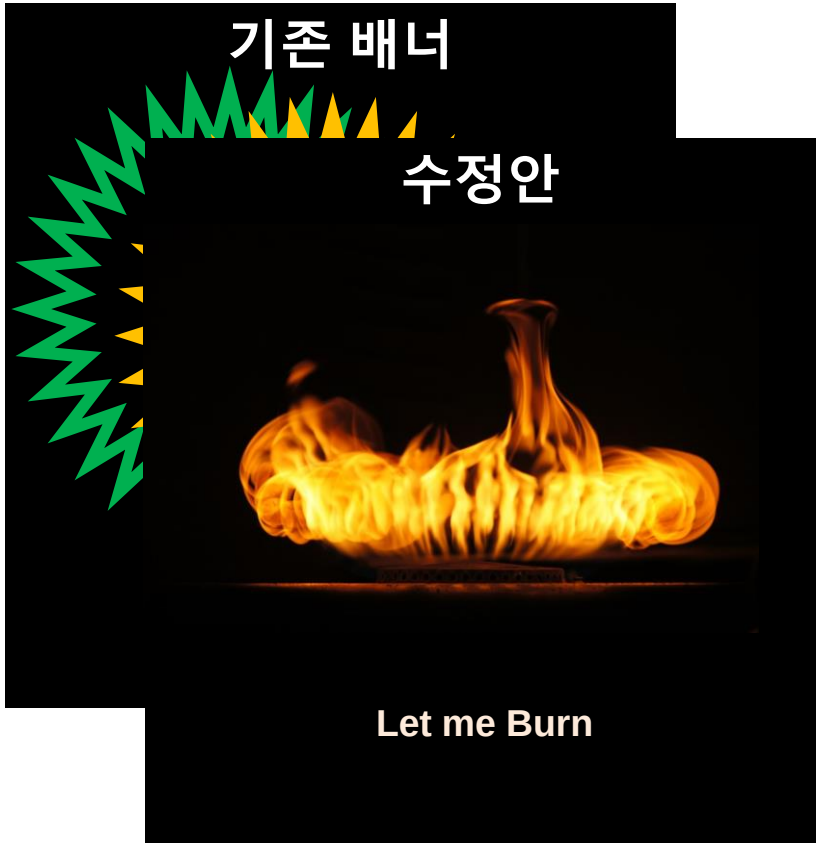
$$= \frac{[\mu_{new} - \mu_{old}]_{from\ Data}}{SE(new, old)} = 3.7442$$

자유도와 표본표준편차 계산의 디테일을 알고 싶다면, https://support.minitab.com/ko-kr/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/2-sample-t/methods-and-formulas/methods-and-formulas/

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

T-test!



귀무 가설, 기존 입장, 보수적인 입장, 현재의 기준

H_0 : 기존 배너든 바꾸든 일 평균 방문수에 차이 없음

H_1 : $\mu_{new} = \mu_{old}$ // $\mu_{new} - \mu_{old} = 0$

대립가설, 새로운 입장

H_0 : 수정안의 일 평균 방문수가 더 나을 거야

H_1 : $\mu_{new} > \mu_{old}$ // $\mu_{new} - \mu_{old} > 0$

관찰할 Signal의 구조 / Signal의 비교 기준

$$t - statistics = \frac{[\mu_{new} - \mu_{old}]_{from\ Data} - [\mu_{new} - \mu_{old}]_{from\ H_0}}{SE(new, old)}$$

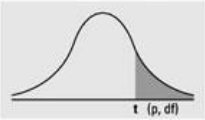
표준편차의 기능 : Noise의 크기 측정

$$\frac{\text{Signal}}{\text{Noise}} = \frac{[\mu_{new} - \mu_{old}]_{from\ Data}}{SE(new, old)} = 3.7442$$

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

귀무 가설, 기존 입장, 보수적인 입장, 현재의 기준

H_0 : 기존 배너든 바꾸든 일 평균 방문수에 차이 없음

H_1 : $\mu_{new} = \mu_{old}$ // $\mu_{new} - \mu_{old} = 0$

대립가설, 새로운 입장

H_0 : 수정안의 일 평균 방문수가 더 나을 거야

H_1 : $\mu_{new} > \mu_{old}$ // $\mu_{new} - \mu_{old} > 0$

관찰할 Signal의 구조 / Signal의 비교 기준

$$t - statistics = \frac{[\mu_{new} - \mu_{old}]_{from\ Data} - [\mu_{new} - \mu_{old}]_{from\ H_0}}{SE(new, old)}$$

표준편차의 기능 : Noise의 크기 측정

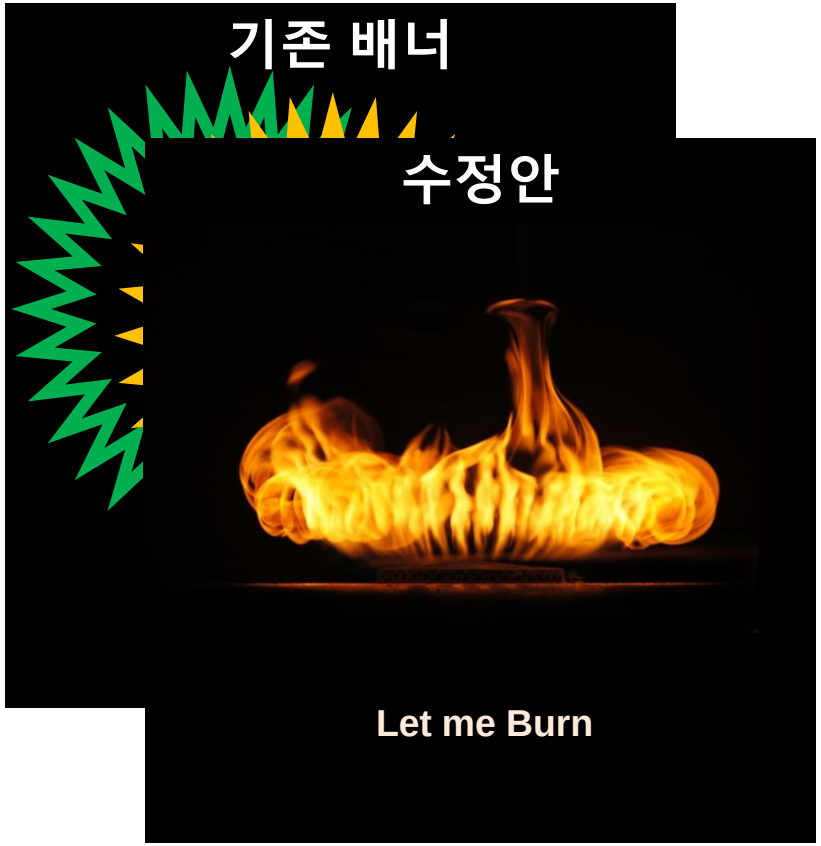
$$\frac{\text{Signal}}{\text{Noise}} = \frac{[\mu_{new} - \mu_{old}]_{from\ Data}}{SE(new, old)} = 3.7442$$

자유도가 22라면?

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

T-test!



P-value가
0.005보다도 작군

통계적으로
유의미함.

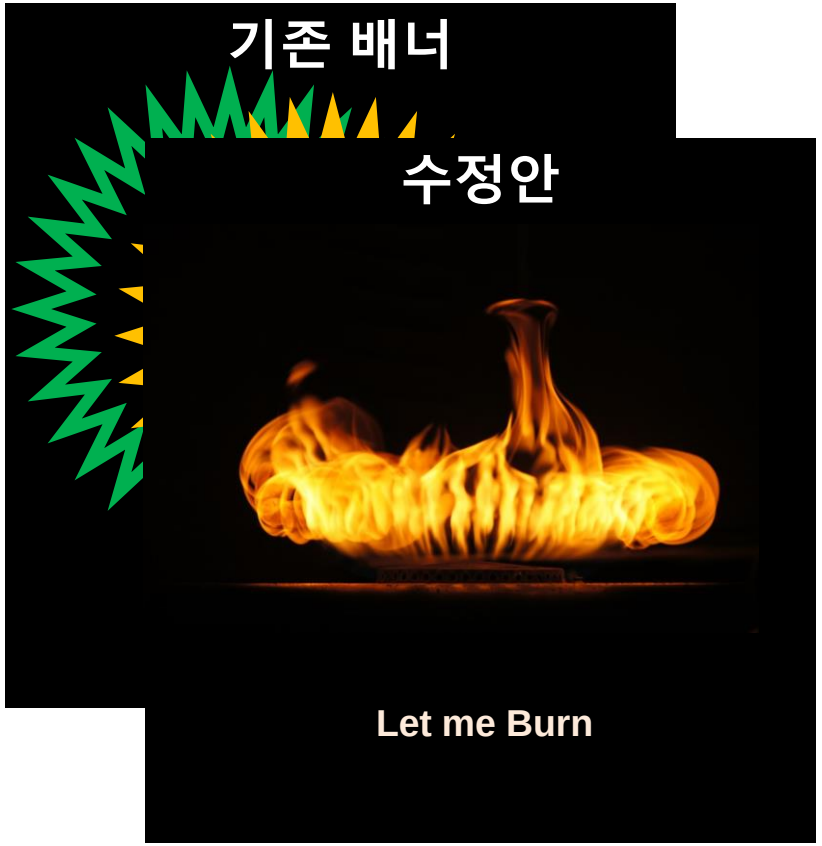
귀무가설이 실험
결과를 설명할 확률이
0.5%도 안 된단
뜻이지

이들의 대화를 한국말로 바꿔볼 것.

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

T-test!



좋아, 차이 자체가
통계적으로 유의미해

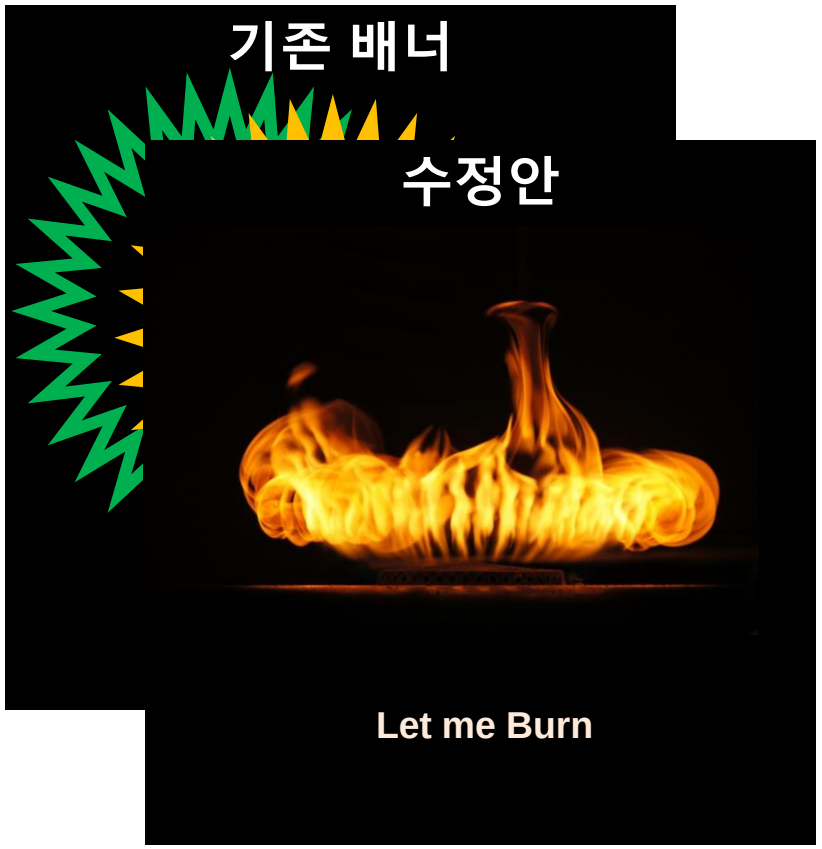
수정안이
기존배너보다는
뛰어나지. 수정안
채택하자!

잠깐, 배너 수정에 드는 비용은 100만원입니다.
기존배너의 일 평균 방문수 300명에
수정안의 일평균 방문수는 330명이지만,
최종적인 Conversion은 20명과 22명 수준으로,
일 평균 10000원 정도 수익의 차이가 납니다.
이벤트는 한달 동안 할건데, 30만원 더 벌자고
100만원을 쓰자구요?

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

T-test!



좋아, 차이 자체가

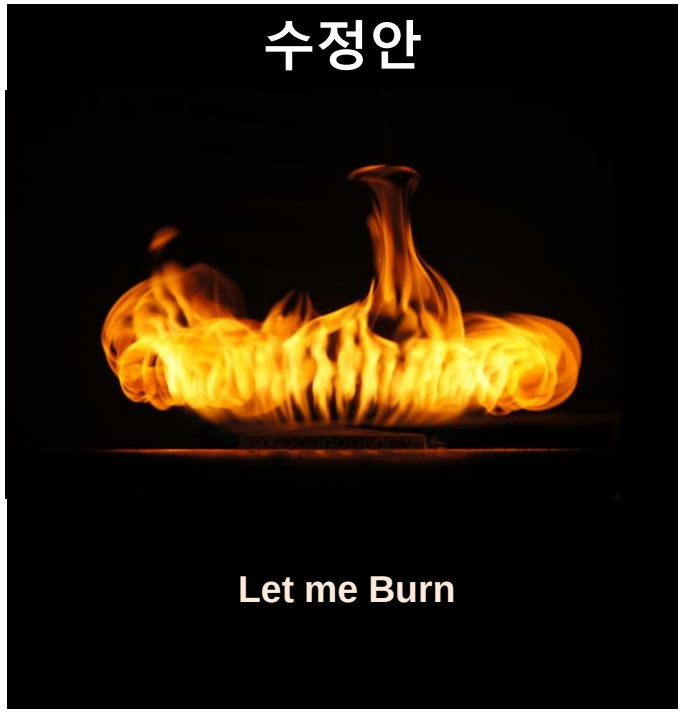
실험 설계 해서 데이터 얻었다고 대뜸
가설검정하지 말고.
데이터를 들여다 봅시다.
현실도 무시해서는 안됩니다.
EDA는 사랑입니다.

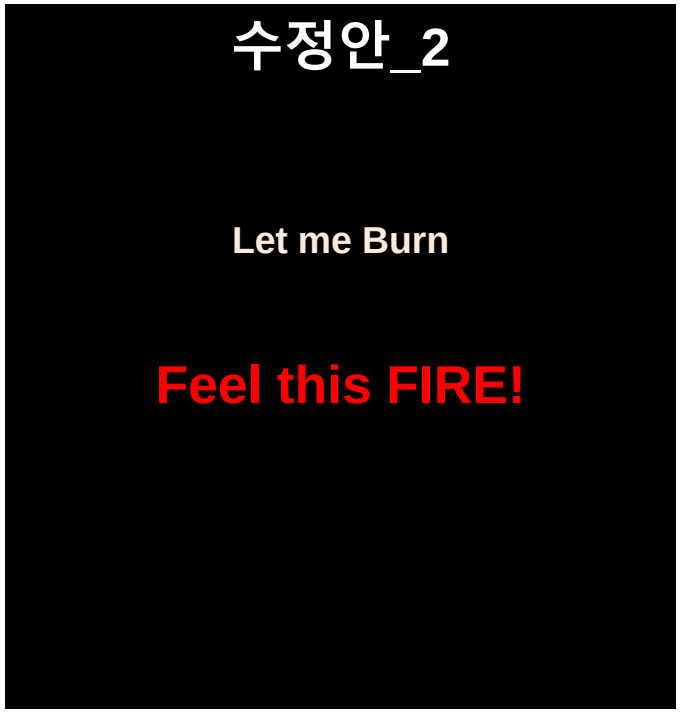
삼간, 배너 수정에 드는 비용은 100만원입니다.
기존배너의 일 평균 방문수 300명에
수정안의 일평균 방문수는 330명이지만,
최종적인 Conversion은 20명과 22명 수준으로,
일 평균 10000원 정도 수익의 차이가 납니다.
이벤트는 한달 동안 할건데, 30만원 더 벌자고
100만원을 쓰자구요?

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

Analysis of Variance : ANOVA





D:plus

4. Comparative Type : T-test부터 ANOVA살짝

Analysis of Variance : ANOVA

가설의 기본 구조와 P-value 이해한 당신,
이제는 검정방법은 그냥 가져다가 써도 좋아요.
문제가 생길 일은 없을 겁니다.

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

Analysis of Variance : ANOVA

가설의 기본 구조와 P-value 이해한 당신,
이제는 검정방법은 그냥 가져다가 써도 좋아요.
문제가 생길 일은 없을 겁니다.

말만 들어도 무서운 ANOVA
비교가 3집단 이상이면 이만큼 편한 방법도 없음.
결국 평균 비교인데, 왜 이름이 분산분석?

살짝 INTUTION을 들여다 봅시다.

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

The Intuition Behind ANOVA

초 간략 판 예시, 각 배너를 타고 들어오는 일 방문객 수

일	기존 배너	New 1안	New 2안
1	49	56	51
2	47	54	55
3	46	61	57
4	50	57	53
평균	48	57	54

원본 데이터
전체 평균 53

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

The Intuition Behind ANOVA

초 간략 판 예시, 각 배너를 타고 들어오는 일 방문객 수

일	기존 배너	New 1안	New 2안
1	49	56	51
2	47	54	55
3	46	61	57
4	50	57	53
평균	48	57	54

원본 데이터
전체 평균 53

일	기존 배너	New 1안	New 2안
1	-4	3	-2
2	-6	1	2
3	-7	8	4
4	-3	4	0

데이터 – 전체 평균

데이터를 **전체 평균으로만 설명하려고 할 때** 생기는 편차 (오차, noise)
총 편차

=

일	기존 배너	New 1안	New 2안
1	1	-1	-3
2	-1	-3	1
3	-2	4	3
4	2	0	-1

데이터 – 그룹 평균

데이터를 **그룹별 평균으로 설명하려고 할 때** 생기는 편차 (오차, noise)
그룹 내 편차

+

일	기존 배너	New 1안	New 2안
1	-5	4	1
2	-5	4	1
3	-5	4	1
4	-5	4	1

그룹평균 - 전체평균

데이터를 **전체 평균으로 설명할 때는 설명이 안되던 오차** 중, **그룹별 평균으로 설명하면 설명이 되는 부분**
그룹 간 편차

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

The Intuition Behind ANOVA

초 간략 판 예시, 각 배너를 타고 들어오는 일 방문객 수

일	기존 배너	New 1안	New 2안
1	49	56	51
2	47	54	55
3	46	61	57
4	50	57	53
평균	48	57	54

원본 데이터
전체 평균 53

일	기존 배너	New 1안	New 2안
1	-4	3	-2
2	-6	1	2
3	-7	8	4
4	-3	4	0

데이터 – 전체 평균

=

일	기존 배너	New 1안	New 2안
1	1	-1	-3
2	-1	-3	1
3	-2	4	3
4	2	0	-1

데이터 – 그룹 평균

+

일	기존 배너	New 1안	New 2안
1	-5	4	1
2	-5	4	1
3	-5	4	1
4	-5	4	1

그룹평균 - 전체평균

전체 평균으로는 이해 못할 오차

그룹별 평균으로는 이해 못할 오차

전체 평균 대비, 그룹별 평균을 사용하면 설명이 되는 오차

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

The Intuition Behind ANOVA

초 간략 판 예시, 각 배너를 타고 들어오는 일 방문객 수

일	기존 배너	New 1안	New 2안
1	49	56	51
2	47	54	55
3	46	61	57
4	50	57	53
평균			

원본 데이터
전체 평균 53

일	기존 배너	New 1안	New 2안
1	-4	3	-2
2	-6	1	1
3	-7	8	4
4	-3	4	0

61이라는 데이터를
전체 평균 53으로 설명하면
8이라는 오차가 생기는데

데이터 – 전체 평균

전체의
NOISE

=

일	기존 배너	New 1안	New 2안
1	1	1	-1
2	-1	0	1
3	3	4	3
4	2	0	-1

61이라는 데이터를
그룹 평균 57로 설명하면
오차가 4

데이터 – 그룹 평균

여전히 놓치는
NOISE

+

일	기존 배너	New 1안	New 2안
1	-3	2	-1
2	-5	0	0
3	-5	4	1
4	-5	4	1

전체 평균으로 설명할
때보다, 설명할 수 있는
오차가 4가 생김!

그룹평균 - 전체평균

캐치할 수 있게 된
NOISE

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

The Intuition Behind ANOVA

초 간략 판 예시, 각 배너를 타고 들어오는 일 방문객 수

단순히 편차로만 계산해서 하나하나 따지고 들면 어색한 부분은 있음.

일	기존 배너	New 1안	New 2안
1	49	56	51
2	47	54	55
3	46	61	57
4	50	57	53
평균	48	57	54

원본 데이터
전체 평균 53

일	기존 배너	New 1안	New 2안
1	-4	3	-2
2	-6	1	2
3	-7	8	4
4	-3	4	0

데이터 – 전체 평균

전체의 NOISE

=

일	기존 배너	New 1안	New 2안
1	1	-1	-3
2	-1	-3	1
3	-2	4	3
4	2	0	-1

데이터 – 그룹 평균

여전히 놓치는 NOISE

+

일	기존 배너	New 1안	New 2안
1	-5	4	1
2	-5	4	1
3	-5	4	1
4	-5	4	1

그룹평균 - 전체평균

캐치할 수 있게 된 NOISE

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

The Intuition Behind ANOVA

이 편차들을, 한번에 요약해서 비교할 필요가 있다.

일	기존 배너	New 1안	New 2안
1	49	56	51
2	47	54	55
3	46	61	57
4	50	57	53
평균	48	57	54

원본 데이터
전체 평균 53

일	기존 배너	New 1안	New 2안
1	-4	3	-2
2	-6	1	2
3	-7	8	4
4	-3	4	0

데이터 – 전체 평균

전체의 NOISE (어쨌든 편차)

=

일	기존 배너	New 1안	New 2안
1	1	-1	-3
2	-1	-3	1
3	-2	4	3
4	2	0	-1

데이터 – 그룹 평균

여전히 놓치는 NOISE (얘도 편차)

+

일	기존 배너	New 1안	New 2안
1	-5	4	1
2	-5	4	1
3	-5	4	1
4	-5	4	1

그룹평균 - 전체평균

캐치할 수 있게 된 NOISE (결국 이것도 편차)

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

The Intuition Behind ANOVA

이 편차들을, 한번에 요약해서 비교할 필요가 있다.

일	기존 배너	New 1안	New 2안
1	49	56	51
2	47	54	55
3	46	61	57
4	50	57	53
평균	48	57	54

원본 데이터
전체 평균 53

일	기존 배너	New 1안	New 2안
1	-4	3	-2
2	-6	1	2
3	-7	8	4
4	-3	4	0

데이터 – 전체 평균

=

일	기존 배너	New 1안	New 2안
1	1	-1	-3
2	-1	-3	1
3	-2	4	3
4	2	0	-1

데이터 – 그룹 평균

+

일	기존 배너	New 1안	New 2안
1	-5	4	1
2	-5	4	1
3	-5	4	1
4	-5	4	1

그룹평균 - 전체평균

어쨌든 편차니까
제곱해서 평균내면 분산!

얘도 결국 제곱해서
평균내면 분산!

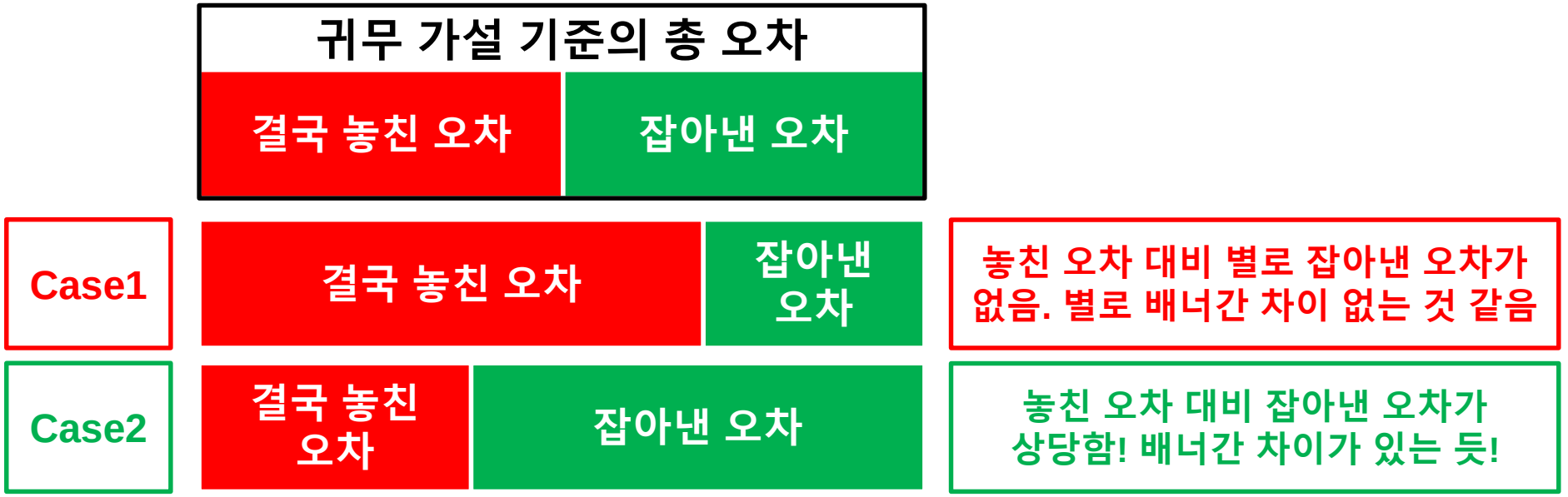
D:plus

4. Comparative Type : T-test부터 ANOVA살짝

The Intuition Behind ANOVA

일	기존 배너	New 1안	New 2안
1	49	56	51
2	47	54	55
3	46	61	57
4	50	57	53
평균	48	57	54

원본 데이터
전체 평균 53



오차의 크기들을 비교해야 하는데 그 요약 수단으로 분산을 사용했음.

그리고 구한 분산을 비교해서 잡아낸 오차가 어느 정도로 큰지 관찰함

그래서 분산분석 이라고 함!

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

즉, 기준(전체 평균 : 귀무 가설 / 그룹 별 평균 : 대립 가설)을
무엇으로 하느냐에 따라서 생기는 오차의 크기를 분산을 이용해서 비교한 것!

귀무 가설 : 모든 그룹 간 평균 차이가 없다 ; 배너에 따른 방문객 수 차이는 없을 것이다.
대립 가설 : 최소한 한 그룹은 평균 차이가 있을 것이다 ; 배너에 따른 방문객 수 차이는 있을 것이다.

SS = Sum of Squares : 제곱해서 더한 거.

MS = Mean of Squares : 제곱해서 더한 거를 자유도로 나누어 평균낸거

변동 요인	변동합(SS)	자유도	분산(MS)	F ratio	P-value	5%
그룹 간 (Between Groups)	SSB : 168	2 (그룹 수 – 1)	MSB : 84	**13.5**	0.002	$F(2,9) = 4.26$
그룹 내 (Within Groups)	SSW : 56	9 (데이터 수 – 그룹 수)	MSW : 6.22			
Total	SST : 224	11 (데이터 수 -1)	(보통 안 구함)			

$$F = \frac{MSB}{MSW}$$

분산 비교 부분!

자유도는 일단, 참자.

Intution에 의한 모델링이라기 보다,
수학적인 보정의 성격이 훨씬 크다.

거칠게 말하면,
놓친 오차보다 잡아낸 오차가
13.5배는 크다!
라고 설명해 볼 수 있음.

D:plus

4. Comparative Type : T-test부터 ANOVA살짝

Summary 아닌 Summary

		Dependent Variable	
		Categorical	Continuous
Independent Variable	Categorical	Chi-squared test	ANOVA
	Continuous	Logistic Regression	Linear Regression

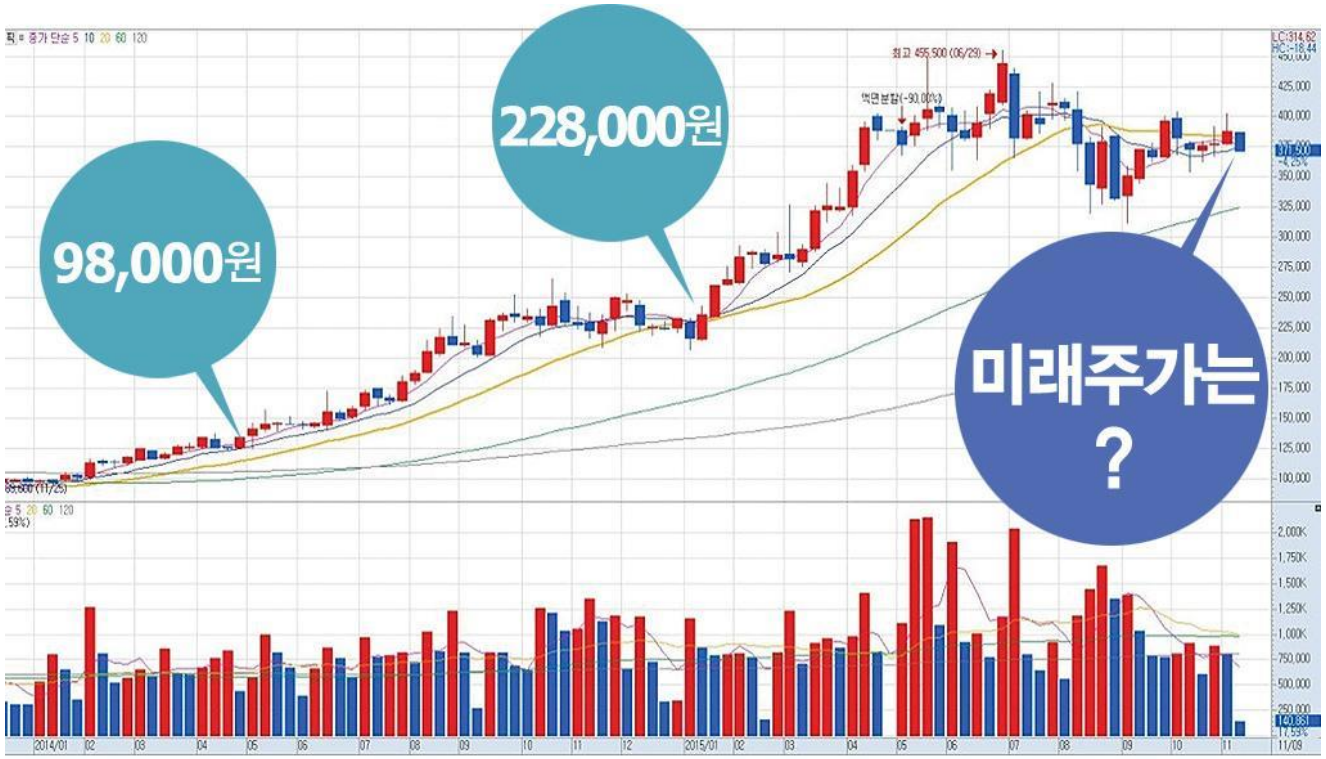
D:plus

5. Predictive type :

Before Machine Learning

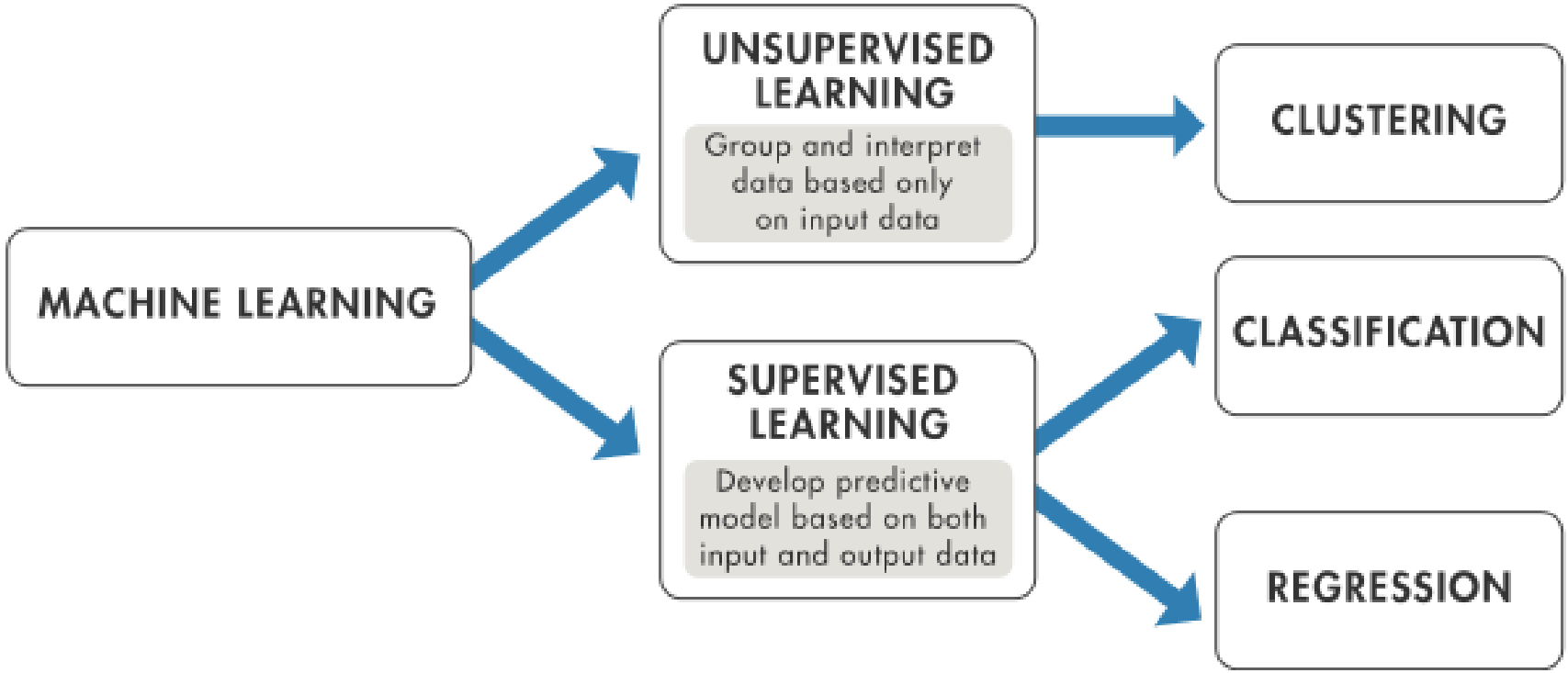
5. Predictive type : Before Machine Learning

이것도 쉽다. 우리의 욕망!



5. Predictive type : Before Machine Learning

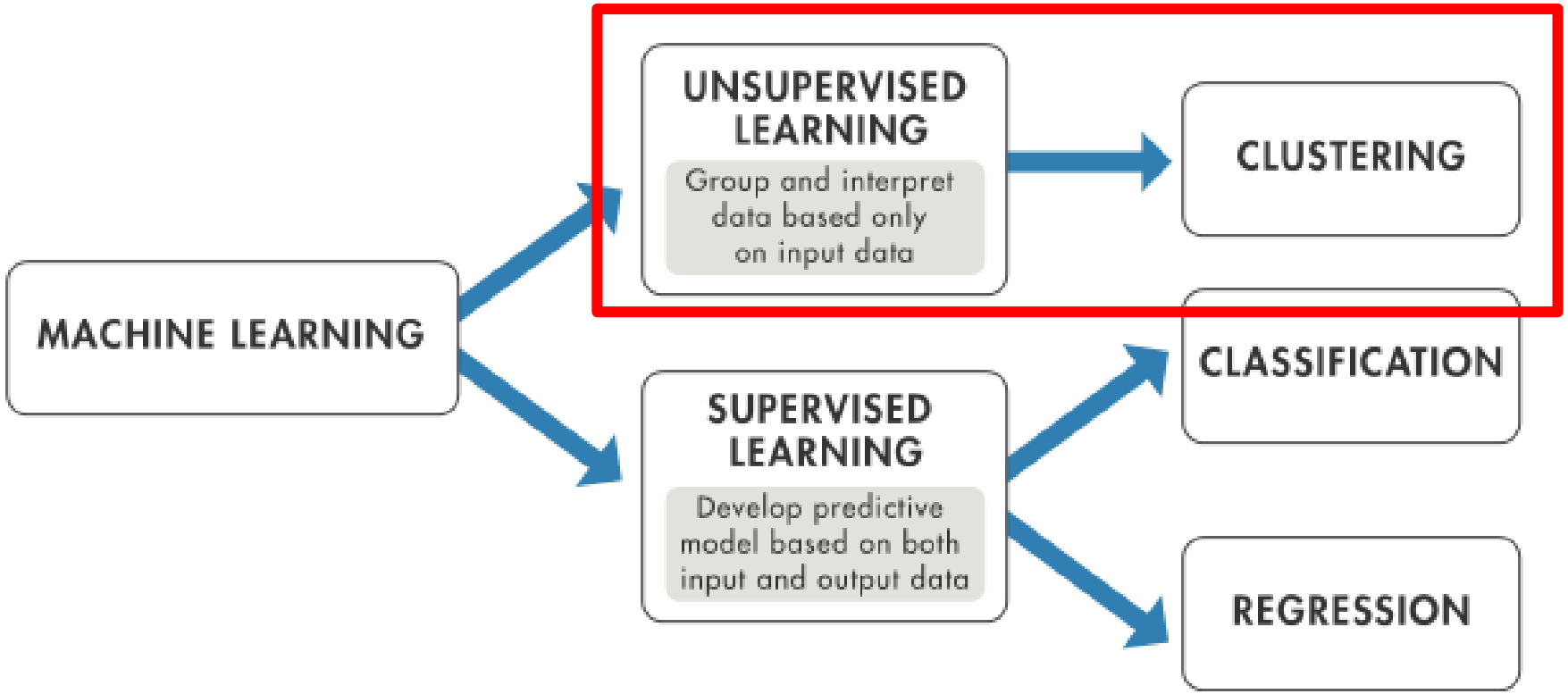
머신러닝은, 앞의 고민들이 끝난 후 배워야 맞다.



D:plus

5. Predictive type : Before Machine Learning

클러스터링을 먼저 배우기 보다는.....

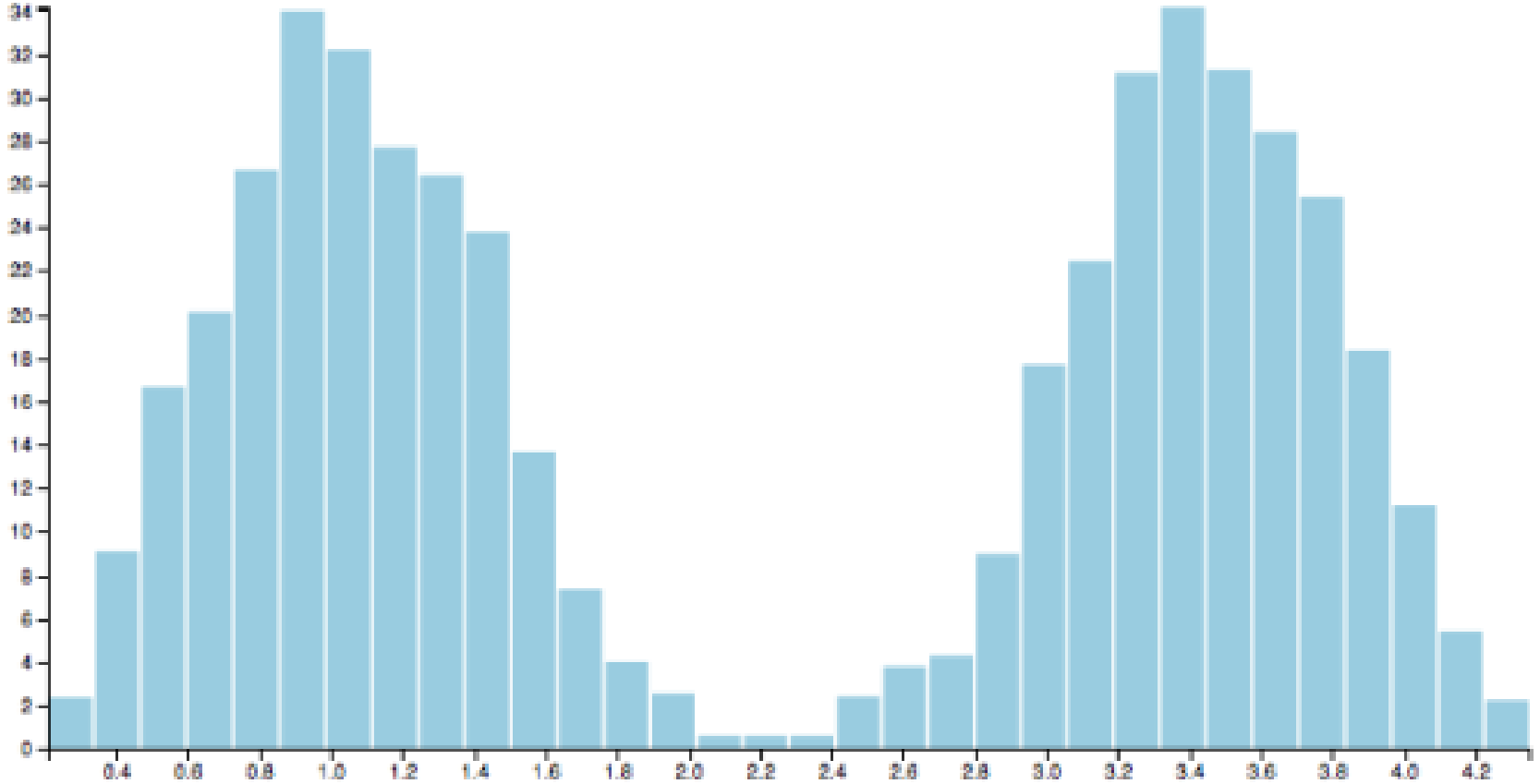


무언가를 Associative하게 Describe하려는 처절한 노력!

D:plus

5. Predictive type : Before Machine Learning

클러스터링이 필요한 상황을 먼저 인지하자.



D:plus

5. Predictive type : Before Machine Learning

클러스터링이 필요한 상황을 먼저 인지하자.

분리 시켜 봐야 합당하거나.
분리 시켜봐도 괜찮을까? 하는 기대로부터 출발한다.

5. Predictive type : Before Machine Learning

클러스터링이 필요한 상황을 먼저 인지하자.

분리 시켜 봐야 합당하거나.
분리 시켜봐도 괜찮을까? 하는 기대로부터 출발한다.
분리시켜서, 군집별로 특징들이 차이가 난다면

새로 수집된 데이터가 어떤 군집일지 예측해보는 것은 의미가 있다.

D:plus

5. Predictive type : Before Machine Learning

클러스터링이 필요한 상황을 먼저 인지하자.

분리 시켜 봐야 합당하거나.
분리 시켜봐도 괜찮을까? 하는 기대로부터 출발한다.
분리시켜서, 군집별로 특징들이 차이가 난다면

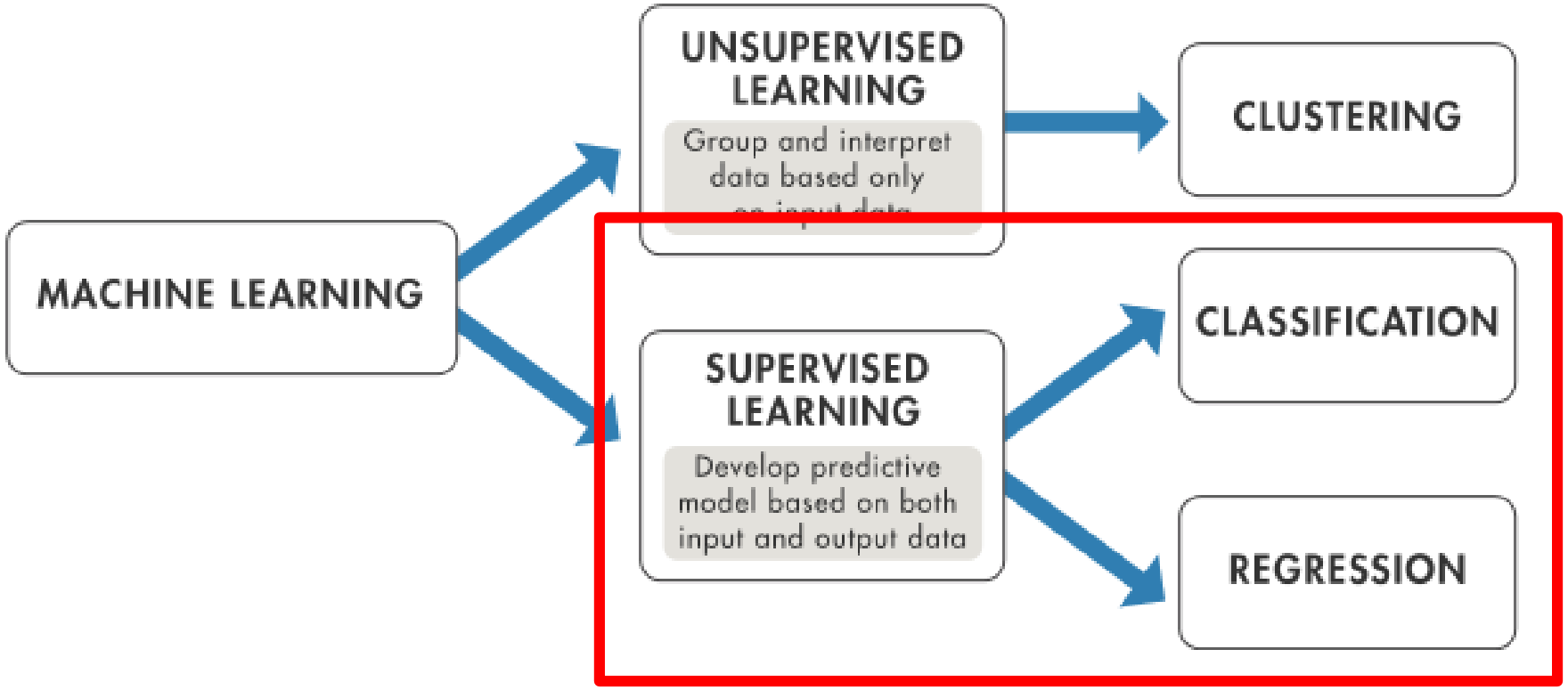
특징들의 차이가 전략으로 이어지는 근거가 된다면!

새로 수집된 데이터가 어떤 군집일지 예측해보는 것은 의미가 있다.

D:plus

5. Predictive type : Before Machine Learning

Supervised Learning은 풀어야 할 문제 / 의문 / 아이디어로 출발해야 한다.



D:plus

5. Predictive type : Before Machine Learning

다시 한번, 에이드 판매!

앞의 세가지 타입의 질문들은 결국 크게 두 가지.

1. 관심사를 주어진 조건들로 부터 이해해보는 것.
2. 선택을 위해 비교해보는 것.

Regression

판매량을 예측가능 하다면 어떤 문제들이 해결이 될까?

D:plus

5. Predictive type : Before Machine Learning

혹은, 난 슈퍼마켓 점장!

앞의 세가지 타입의 질문들은 결국 크게 두 가지.

1. 관심사를 주어진 조건들로 부터 이해해보는 것.
2. 선택을 위해 비교해보는 것.

Classification

이탈할 사람들을 미리 알 수 있다면 무슨 문제를 해결할 수 있을까

D:plus

5. Predictive type : Before Machine Learning

아래의 질문에 답을 할 수 있어야 한다.

현실을 무시해서는 안 된다.

Regression

판매량을 예측가능 하다면 어떤 문제들이 해결이 될까?

Classification

이탈할 사람들을 미리 알 수 있다면 무슨 문제를 해결할 수 있을까

D:plus

5. Predictive type : Before Machine Learning

Summary 아닌 Summary

그리고, 바로 머신러닝으로 넘어가는 것이 아니라!

앞에서 진행한 분석 + Alpha를 이용해 과거를 충분히 설명해 봐야 한다.

1. **내가 캐치하고자 하는 Signal의 구조를 충분히 확인할 것.**
   → 관심사를 어떻게 설명해낼지 충분히 구조를 잡아내야 함.

2. **그 구조를 잡고, 미래의 데이터에 대해서 확인해보면 됨.**
   Case1 → 내가 가진 데이터가 미래에 얻을 데이터에 대해서도 대표성이 있다면, 미래 예측은 성공할 수 있음.
   Case2 → 미래에 얻게 될 데이터가 과거의 데이터와 확 다르면 애당초 가능성이 없음.