The background features a dark, abstract design with a glowing white hexagon at the top center, connected by thin white lines to other points. Below this, there are several overlapping, glowing white geometric shapes, including a large '10' and various polygons. The overall aesthetic is futuristic and data-oriented.

빅데이터의 이해와 활용

Understanding and Using Big Data

10

통계분석

학습 내용

- 01 가설검정
- 02 상관분석
- 03 회귀분석

학습 목표

- 가설 검정의 절차, 유의수준, 기각역 등에 대한 개념을 이해하고 가설 결과를 기술할 수 있다.
- 상관분석을 이해하고 상관계수와 그래프를 보고 상관성을 설명할 수 있다.
- 회귀분석에 대해 이해하고 독립변수와 종속변수와의 관계에 대해 설명할 수 있다.

생각 해보기

작년 통계청에서 조사한 만 7세 남자 어린이의 키의 평균이 1220mm라고 한다.

올해 우리 아들이 7살이 되었는데 키가 1330mm이다.
우리 아들의 키는 7세 남자 아이의 평균일까?
아니면 어떠한 수준일까?



01

가설검정

- 1) 가설검정이란
- 2) 가설검정 과정
- 3) 가설 수립
- 4) 검정 통계량
- 5) 가설검정의 종류
- 6) 검정 통계량 계산
- 7) 유의수준과 기각역
- 8) 판정
- 9) t-분포표
- 10) 검정 결과 기술

1) 가설검정이란

가설검정(Hypothesis Test)

모수의 상태에 대한 여러 주장들 중 어떤 주장을 사실로
받아들일지를 결정하는 과정



1) 가설검정이란

● 예



작년도 대한민국 ‘만7세 남자 어린이의 키의 평균이 1220mm’라는 기준에 알려진 모수의 상태가 올해에도 여전히 유의한지 확인하는 방법

- 모수의 참값을 구하는 것이 아니라 모수의 상태가 ‘만7세 남자 어린이의 키의 평균이 1220mm’라는 기준에 알려진 사실이 현재에도 유지되고 있는지 확인하는 것
- 만 7세 남자 어린이들의 모집단에서 15명의 어린이를 표본으로 추출하여 키를 조사함

1196	1340	1232	1184	1295
1247	1201	1182	1192	1287
1159	1160	1243	1264	1276

2) 가설검정 과정

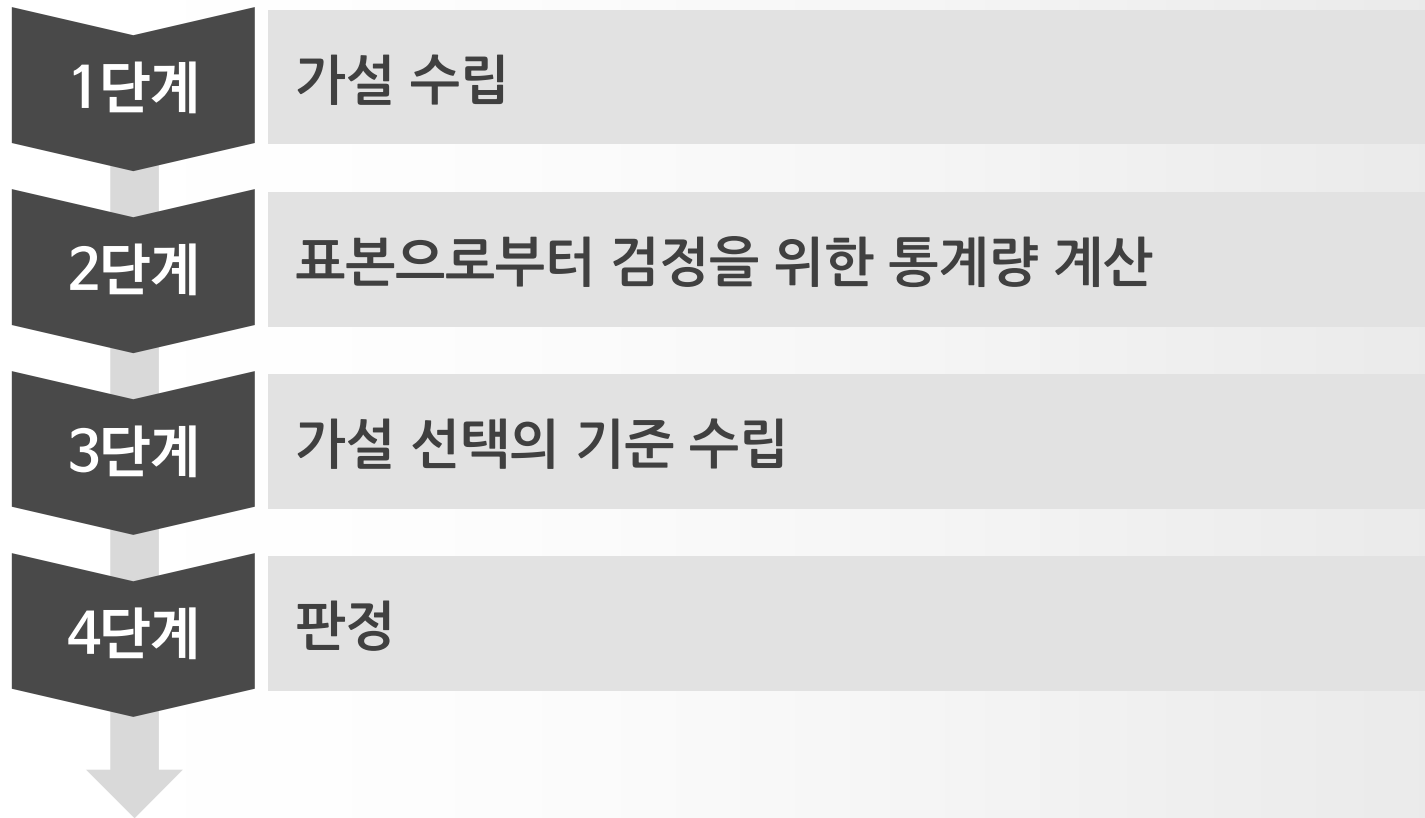
가설검정 과정

모집단 특성의 상태에 대한 주장인 가설에 대해
표본으로부터 얻은 정보를 바탕으로
이를 채택할지 기각할지를 판단함으로써 모집단의 상태에
대해 결정하는 과정



2) 가설검정 과정

“ 다음의 4단계를 거쳐 이뤄짐 ”



3) 가설 수립

- 가설검정에서 사용하는 가설의 종류

영가설 (귀무가설, Null Hypothesis, H_0)

주로 기존에 알려진 것과 차이가 없음을 나타내는 가설

대안가설 (대립가설, Alternative Hypothesis, H_1)

주로 기존에 알려진 것과 차이가 있음을 나타내는 가설

↳ 연구자가 밝히고자 하는 가설로 연구 가설이라고도 함

3) 가설 수립

- 가설검정에서 사용하는 가설의 종류

- ▶ 영가설과 대안가설 수립의 예

가설	내용	수식표현
영가설 H_0	(만7세 남자 어린이의) 키의 평균은 1220mm이다.	$\mu_{키} = 1220(mm)$
대안가설 H_1	(만7세 남자 어린이의) 키의 평균은 1220mm가 아니다.	$\mu_{키} \neq 1220(mm)$

4) 검정통계량

검정 통계량(Test Statistics)

영가설의 채택 및 기각 여부를 확인하기 위해 표본을 통해 관찰된 값을 사용하는 통계량



4) 검정통계량



검정 통계량의 계산은 표본으로부터 관찰된 특성



모수의 상태로 '영가설이 참'이라는 가정 하에 계산하고,
판정단계에서 이 가정을 유지할 것인지의 여부를 결정함

- 영가설이 참이라는 가정을 받아들일 수 없을 때
영가설을 기각함
- 영가설이 참이라는 가정을 받아들일 때는 영가설을
채택함

5) 가설검정의 종류

표본의 개수	검정대상	모분산 파악여부	분석 구분
1개	평균	알고 있음(Known)	한 표본에서 평균에 대한 Z-검정
		모름(Unknown)	한 표본에서 평균에 대한 t-검정
	비율	관계없음	한 표본에서 비율에 대한 비율 검정
	분산	관계없음	한 표본에서 분산에 대한 모분산 검정
2개	평균	관계없음 / 독립된 표본	두 표본에서 평균에 대한 t-검정
		관계없음 / 쌍체 표본	두 표본에서 평균에 대한 쌍체 t-검정
	비율	관계없음	두 표본에서 비율에 대한 비율 검정
	분산	관계없음	두 표본에서 분산에 대한 모분산 검정

6) 검정 통계량 계산

- 검정통계량 계산의 예

- ▶ 만 7세 남자 어린이의 평균 키에 대한 가설검정

모집단

‘만7세 남자 어린이의 키’ 하나이고,
평균에 대한 가설검정을 하는 경우

➡ 모집단이 한 개일 경우의 평균 검정



한 개의 모집단 특성의 평균에 대한 검정에서는
모집단의 분산을 모를 때 t-통계량을 사용함



$$T = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} \sim t(n - 1)$$

6) 검정 통계량 계산

- 검정통계량 계산의 예

- ▶ 만 7세 남자 어린이의 평균 키에 대한 가설검정

검정통계량 계산

- \bar{X} 는 표본평균, s 는 표본표준편차, n 은 표본의 개수로 표본으로부터 계산함
- μ_0 는 우리가 알고자 하는 모평균으로, 모집단의 평균 1220mm임
- 표본으로부터 구한 검정통계량은 ‘영가설이 참’일 때 자유도가 $n-1$ 인 t -분포에서 관찰된 값임

6) 검정 통계량 계산

- 검정통계량 계산의 예

- ▶ 만 7세 남자 어린이의 평균 키에 대한 가설검정

관찰된 자료의 평균	1230.533(mm)
영가설 하에서 모평균	1220(mm)
표본의 표준편차	54.186(mm)
표본의 크기	15
자유도	14

6) 검정 통계량 계산

- 검정통계량 계산의 예

- ▶ 만 7세 남자 어린이의 평균 키에 대한 가설검정



모분산을 알지 못하기에 t-분포를 가정하고
t-분포 통계량을 구하면 약 0.727임







$$T = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} = \frac{1230.533 - 1220}{54.186 / \sqrt{15}} \cong 0.727$$

7) 유의수준과 기각역

- 가설검정 시(판정의) 오류

{ 실제 영가설이 참일 때 가설검정을 통해 대안가설을 선택하거나,
실제 영가설이 거짓일 때 가설검정을 통해 영가설을 선택하는 경우 }

	Reality	
	True(H_0)	False(H_1)
True(H_0)		 (Type1 Error, α)
False(H_1)	 (Type2 Error, β)	
Measured		

7) 유의수준과 기각역

- 가설검정 시(판정의) 오류

제 1종 오류

영가설이 참인데 대안가설을
선택하는 오류

제 2종 오류

영가설이 거짓인데 영가설을
선택하는 오류



7) 유의수준과 기각역

어떤 오류를 관리할 것인가?

법정에서는 국민참여재판을 신청한 피고인 A씨에 대한 법정공방이 벌어지고 있습니다.

변호인은 피고인이 무죄라고 주장하면서, 판사와 배심원단을 상대로 설전을 벌이고 있습니다.

각종 증거와 반론이 오고 간 법정공방을 마치고, 배심원단은 숙고 끝에 평결을 판사에게 전달하고 판사는 배심원단의 의견을 참고하여 판결을 내리고자 합니다.



7) 유의수준과 기각역

- 어떤 오류를 관리할 것인가?

[판사는 A씨의 범죄에 대해 무죄 혹은 유죄를 판결함에 있어, 유죄인 상태를 영가설, 무죄인 상태를 대안가설이라고 하면 잘못된 판단은 두 가지가 있음]

01

실제 무죄이나 유죄 판결을 받아 양형에 따른 수감 생활(제2종 오류)

02

실제 유죄이지만 무죄 판결을 받아 사회로 돌아감 (제1종 오류)

7) 유의수준과 기각역

● 어떤 오류를 관리할 것인가?

여러 사람이 함께 어울리는 사회의 관점에서 볼 때

- ①의 무죄이나 억울한 옥살이를 하게 되는 경우도 문제가 됨
- 사회적인 관점에서만 보자면 ②의 죄인이 죄값을 치루지 않고 유유히 법정을 나와 사회의 구성원이 되는 것이 더 큰 문제가 됨



제1종 오류의 경우, 영가설이 참이지만 참이 아니라고 주장하는 경우임

- 연구에서는 차이가 없으나 차이가 있다고 주장하는 경우로 제1종 오류가 더 심각한 상황이 될 것임

7) 유의수준과 기각역

- 유의수준(α)

유의수준(α)

제 1종 오류를 범할 확률의 최대 허용 한계를
유의수준이라고 함

└ 연구에 따라 0.1, 0.05, 0.01 등 여러 기준이 있으나,
통상적으로 유의수준 0.05를 많이 사용함

7) 유의수준과 기각역

- 유의수준의 역할 : 기각역 수립

유의수준(α)

오류가 발생할 확률로써 이는 영가설 하에서 생성되는
표본분포에서의 확률



7) 유의수준과 기각역

- 유의수준의 역할 : 기각역 수립

▶ 예 : 만 7세 남자 어린이의 평균 키에 대한 가설검정에서의 유의수준 정하기

대안가설
(H_1 , Alternative Hypothesis)

(만 7세 남자 어린이의)
키의 평균은 1220mm가 아니다. ($\mu_7 \neq 1220$)



이 대안가설을 만족하는 상황은 두 가지임

- 검정통계량 T 가 1220mm보다 현저히 작은 경우 ($T < c_l$)
- 검정통계량 T 가 1220mm보다 현저히 큰 경우 ($T > c_u$)



‘~보다 크다’, ‘~보다 작다’는 상대적인 개념으로 기준이 되는 값이 필요하고
유의수준이 그 기준을 제시해주는 역할을 함

7) 유의수준과 기각역

- 유의수준을 이용한 기준 제시

유의수준을 이용한 기준 제시

- 예에서는 대안가설에 의해 작은 쪽과 큰 쪽 두 곳의 기준이

- 1 작은 쪽의 기준을 c_l 이라 할 때, c_l 은 영가설 하의 분포에서 $P(T < c_l) = \alpha/2$ 가 되게 하는 값
- 2 큰 쪽의 기준을 c_u 이라 할 때, c_u 은 영가설 하의 분포에서 $P(T > c_u) = \alpha/2$ 가 되게 하는 값

7) 유의수준과 기각역

- 유의수준을 이용한 기준 제시

유의수준을 이용한 기준 제시

- 예에서는 대안가설에 의해 작은 쪽과 큰 쪽 두 곳의 기준이

3 이 기준에 따라 $\alpha = 0.05$ 라 했을 때 c_l 보다 작은 쪽의 확률이 0.025가 되게 하는 영가설 하에서 값을 R을 이용해 계산함

$$> qt(0.025, df = 14) \\ [1] - 2.144787$$

➡ $P(T < c_l) = 0.025$ 인 자유도가 14인 t-분포에서의 값 c_l 은 약 -2.14이며
 $P(T > c_u) = 0.025$ 인 자유도가 14인 t-분포에서의 값 c_u 는 t-분포의 좌우대칭을
이용하여 약 2.14임을 알 수 있음

7) 유의수준과 기각역

- 임계값, 기각역과 채택역

임계값

여기서 구한 두 값 $c_l = -2.14$ 와 $c_u = 2.14$

기각역

분포의 중앙을 중심으로 임계값 바깥쪽의 영역
 $T < c_l, T > c_u$

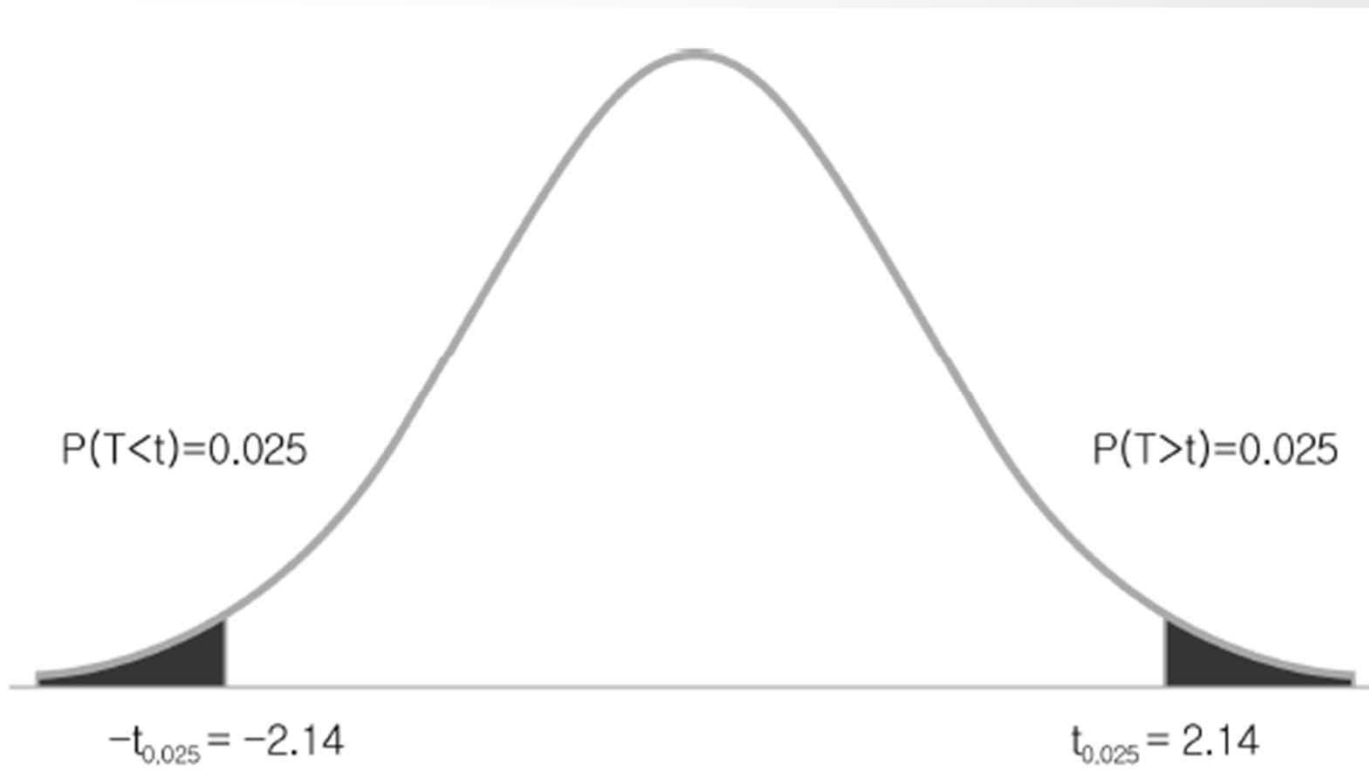
채택역

분포에서 기각역이 아닌 영역 즉, $c_l < T < c_u$

7) 유의수준과 기각역

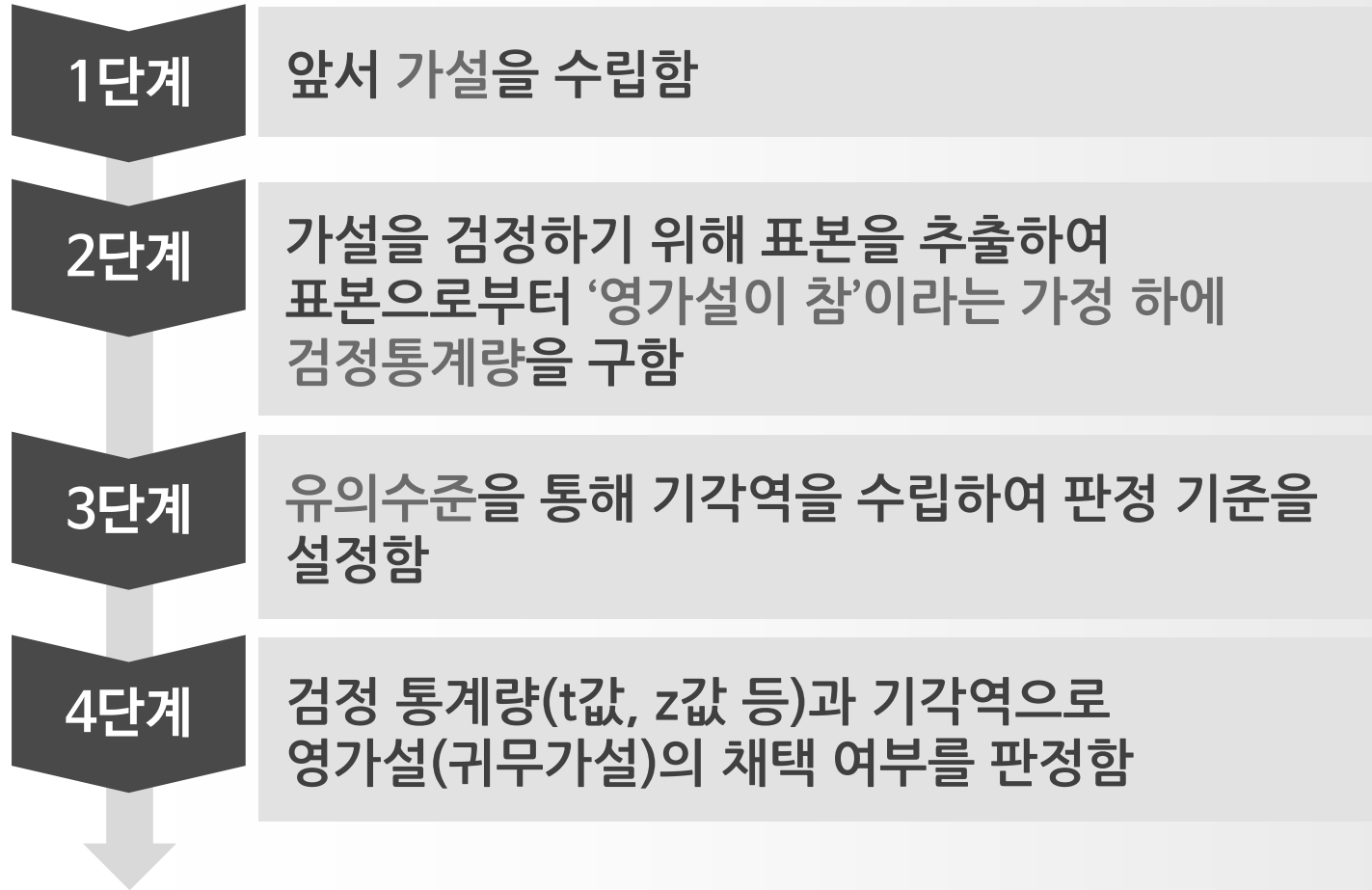
- 임계값, 기각역과 채택역

{ 다음의 그래프에서 붉은 영역이 기각역임 }



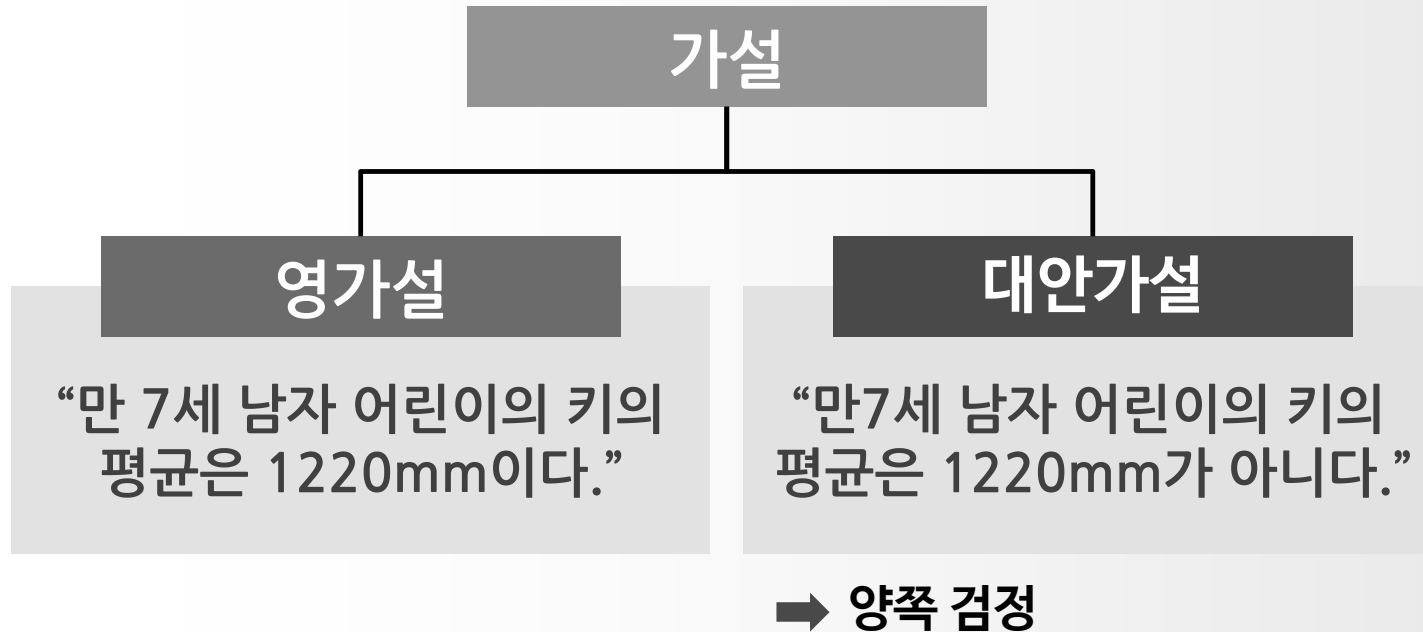
8) 판정

● 판정 단계



8) 판정

- 검정통계량과 기각역을 이용한 판정



8) 판정

- 검정통계량과 기각역을 이용한 판정

검정통계량	0.727(자유도가 14인 t-분포에서)
유의수준	0.05

↳ 기각역은 $T < -2.14, T > 2.14$ 의 두 곳(양쪽 검정)

8) 판정

- 검정통계량과 기각역을 이용한 판정

- ▶ 판정

01

검정통계량이 기각역에 있으면, 영가설을 기각하고
대안가설 채택

02

검정통계량이 기각역에 있지 않으면, 영가설 채택
(대안가설 기각)

- 표본으로부터 구한 검정통계량은 0.753으로
기각역에 존재하지 않아 영가설 채택

➔ 영가설이 참일 때 ($\mu_K = 1220\text{mm}$) 모평균에 대한
표본평균의 분포에서 충분히 발생할 수 있는 경우로
영가설이 참이라는 가정을 뒤집을 만한 근거가 되지 못함

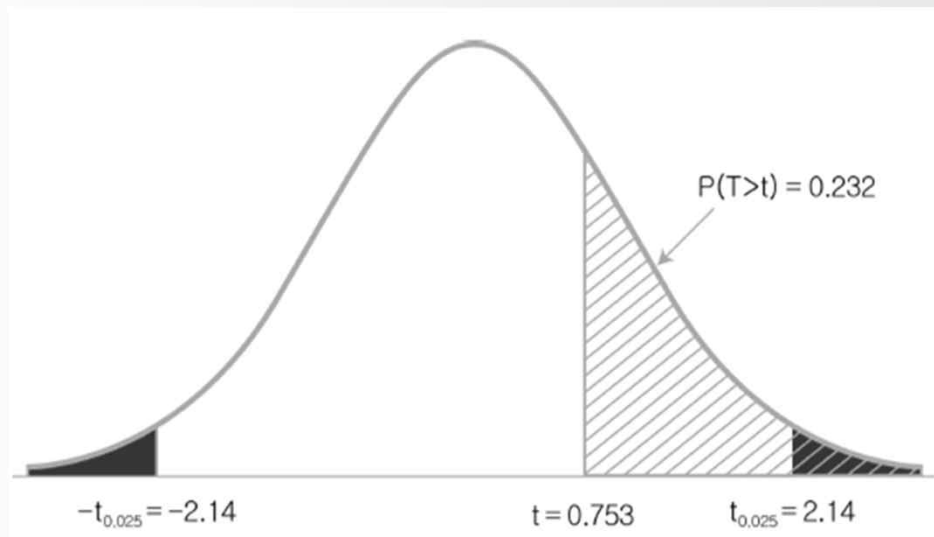
8) 판정

- 앞서 구한 검정통계량에 대한 유의확률 계산 방법



자유도가 14인 t-분포에서 구한 검정통계량 0.753에 대한 유의확률은 $P(T > 0.753)$ 로 R에서 다음과 같이 구할 수 있음

$>1 - pt(0.753, df = 14)$
[1] 0.2319624



▶ t-분포표

α df	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.385	1.845	2.282	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.375	1.833	2.262	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

10) 검정 결과 기술

● 검정 결과 기술하기



판정의 근거와 함께 영가설의 채택 및 기각 여부를
가설검정의 결론으로 표현함



가설검정에서는 판정 과정이 마지막이지만,
판정 결과를 다른 사람들에게 알리기 위한 것

근거를
제시할 때

검정 통계량 계산에 사용된 표본의
특성 및 계산된 검정 통계량과 유의확률을
같이 제시하여 판정에 대한 근거로 활용함

10) 검정 결과 기술

- 검정 결과 기술 형태

- 01 가설검정을 통해 밝히고자 하는 연구의 내용
- 02 표본으로부터 측정된 일반적 특성 및 검정 통계량 계산의 근거가 되는 통계량
- 03 검정 통계량과 유의확률
- 04 판정의 내용
- 05 가설 검정으로부터 알 수 있는 사실

10) 검정 결과 기술

- 검정 결과 기술의 예시

만 7세 어린이의 키의 평균에 대한 가설 검정

- 1 “(만7세 남자 어린이의) 키의 평균이 1220mm”라는
기존 사실이 현재에도 유지되고 있는지 알아보기 위함
- 2 15명의 7세 어린이를 표본으로 추출하여
키를 측정한 결과, 평균은 1230.53(mm),
표준편차는 54.186(mm)이었음
- 3 표본으로부터 구한 검정 통계량은
0.753(유의확률 : 0.232)로 나타났음

10) 검정 결과 기술

- 검정 결과 기술의 예시

만 7세 어린이의 키의 평균에 대한 가설 검정

- 4 이는 유의수준 0.05에서 “(만 7세 남자 어린이의) 키의 평균이 1220mm이다.”라는 영가설을 기각할 수 없는 검정 결과임
- 5 “(만 7세 남자 어린이의) 키의 평균이 1220mm가 아니다.”라는 대안가설에 대해 통계적으로 유의한 결론을 얻을 수 없었으며, (만 7세 남자 어린이의) 키의 평균이 1220mm라는 기존의 사실은 여전히 유지되고 있는 것으로 판단됨



02

상관분석

- 1) 상관분석이란
- 2) 공분산
- 3) 상관계수
- 4) 분석대상 데이터
- 5) 산점도 결과
- 6) 공분산식과 상관계수식
- 7) 산점도 그리기

1) 상관분석이란

상관분석(Correlation Analysis)

두 변수 간에 어떤 선형적 또는 비선형적 관계를 가지고 있는지 분석하는 방법

상관관계
(Correlation
Coefficient)

두 변수는 서로 독립적인 관계이거나
상관된 관계일 수 있으며
이 때 두 변수 간의 관계의 강도

1) 상관분석이란

상관관계의
정도를 나타내는
단위

모상관계수로 ρ (ρ_w 로),
표본 상관 계수로 r 을 사용함

양적 자료

피어슨 상관계수
(Pearson Correlation
Coefficient)

질적 자료

스피어만 상관계수
(Spearman Correlation
Coefficient)

1) 상관분석이란

- 상관 계수 해석 방법

01 r 이 -1.0과 -0.7 사이이면, 강한 음적 선형관계

02 r 이 -0.7과 -0.3 사이이면, 뚜렷한 음적 선형관계

03 r 이 -0.3과 -0.1 사이이면, 약한 음적 선형관계

04 r 이 -0.1과 +0.1 사이이면, 거의 무시될 수 있는 선형관계

1) 상관분석이란

- 상관 계수 해석 방법

05 r 이 +0.1과 +0.3 사이이면, 약한 양적 선형관계

06 r 이 +0.3과 +0.7 사이이면, 뚜렷한 양적 선형관계

07 r 이 +0.7과 +1.0 사이이면, 강한 양적 선형관계

2) 공분산

공분산(Covariance)

두 확률변수 사이의 관계를 선형관계로 나타낼 때
두 변수 사이 상관의 정도를 나타냄

$$\text{Cov}(X, Y)$$

$$= E[(X - E(X))(Y - E(Y))]$$

$$= E[(X - \mu_X)(Y - \mu_Y)], \\ E(X) = \mu_X, E(Y) = \mu_Y$$

2) 공분산



두 확률변수 X, Y 의 공분산은 $Cov(X, Y)$ 로 표기함



공분산이 갖는 값에 따라 두 확률변수의 관계를 확인할 수 있음



2) 공분산

$$\text{Cov}(X, Y) > 0$$

두 확률변수 X, Y 의 변화가 같은 방향임

➡ 즉, X 증가하면 Y 도 증가하고,
반대로 한 변수가 감소하면 같이 감소함

$$\text{Cov}(X, Y) < 0$$

두 확률변수 X, Y 의 변화가
다른 방향임을 나타냄

➡ 즉, X 증가하면 Y 는 감소하고,
반대로 한 변수가 감소하면 다른 하나는 증가함

$$\text{Cov}(X, Y) = 0$$

두 확률변수 간에 어떠한 (선형) 관계가
없음을 나타냄

3) 상관계수

상관계수(Correlation Coefficient)

두 확률변수 X, Y 의 공분산을 각 확률변수의 표준편차의 곱으로 나눈 값을 (모)상관계수라 하고, 기호로 ρ_{XY} (혹은 ρ)로 나타냄

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y}$$

3) 상관계수



상관계수는 -1부터 1사이의 값을 가짐

- 공분산의 경우 자료의 단위에 따라 값의 크기가 일정하지 않아 비교하기 어려움
- 공분산의 성질을 그대로 이어 받아 두 변수 간의 변화의 방향이 같으면 양수, 반대이면 음수로 나타남



상관계수는 모집단의 특성 중에 하나로 일반적으로 알 수 없음

➡ 두 확률변수로부터 추출한 표본의 특성을 통해 구하는 (피어슨의) 표본상관계수를 이용하여 추정함

3) 상관계수

- 표본공분산과 표본상관계수

표본공분산

- 두 확률변수 X, Y 로 부터 추출한 n 개의 표본 쌍 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 에서
- 확률변수 X 로 부터 추출한 표본 x_1, x_2, \dots, x_n 의 평균을 \bar{x} , 표준편차를 s_x ,
- 확률변수 Y 로 부터 추출한 표본 y_1, y_2, \dots, y_n 의 평균을 \bar{y} , 표준편차를 s_y 라 하면,
- 표본공분산 $cov(x, y)$ 는 다음과 같이 두 표본의 편차의 곱을 모두 합하고 이를 자료의 개수(표본 쌍의 개수) -1로 나누어 계산함

$$cov(x, y)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

3) 상관계수

- 표본공분산과 표본상관계수

표본상관계수

- 표본공분산을 각 표본의 표준편차의 곱으로 나누어 계산함

r

$$= \frac{\text{cov}(x, y)}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}}$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

3) 상관계수

- 표본공분산과 표본상관계수

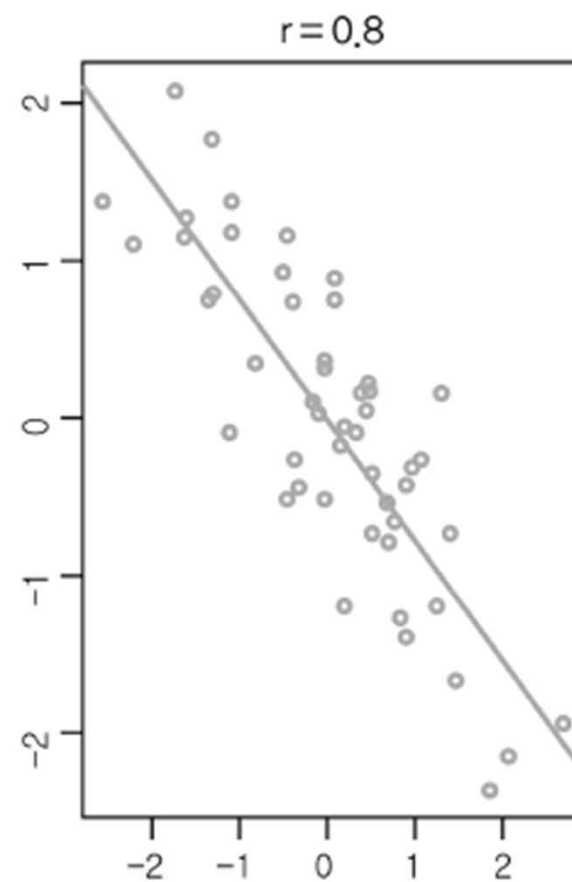
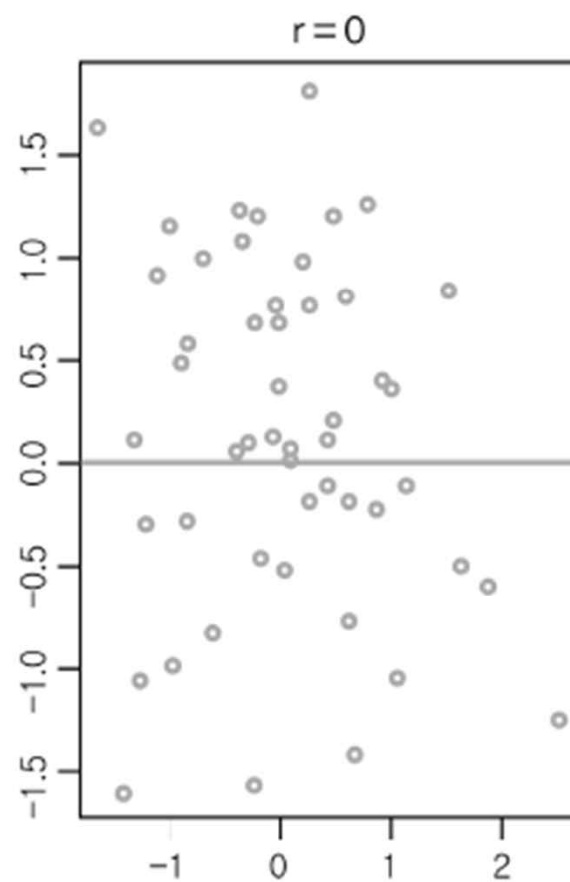
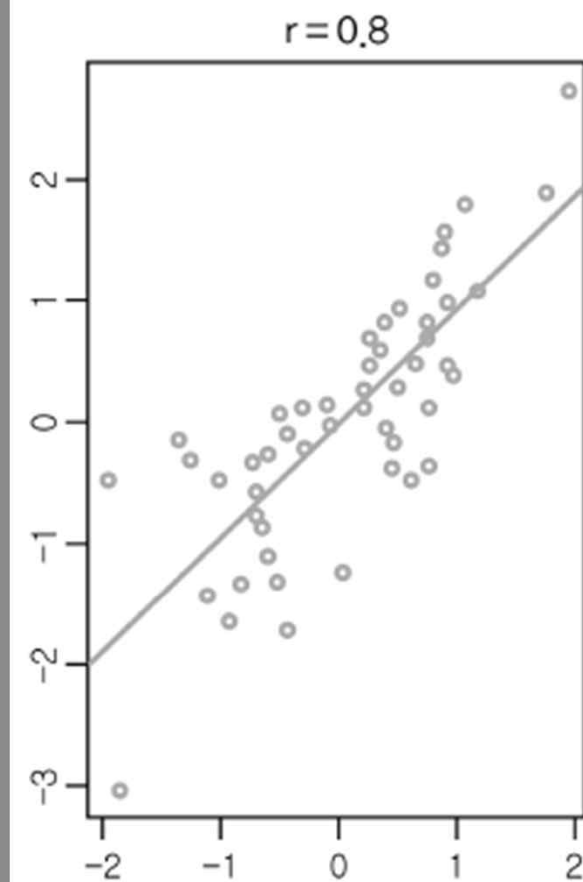
{ 표본상관계수는 모상관계수와 동일한 성질을 가짐 }

01 -1 혹은 1에 가까울수록 강한 상관을 나타냄

02 0에 가까이 갈수록 약한 상관을 나타냄

03 양수일 경우 두 변수의 값의 변화는 같은 방향으로
진행되고, 음수일 경우 값의 변화는 서로 반대가 됨

▶ 표본상관계수



4) 분석대상 데이터

- 데이터 예시

```
> head(hf)
  Family Father Mother Gender Height Kids
1     1    78.5   67.0      M   73.2    4
2     1    78.5   67.0      F   69.2    4
3     1    78.5   67.0      F   69.0    4
4     1    78.5   67.0      F   69.0    4
5     2    75.5   66.5      M   73.5    4
6     2    75.5   66.5      M   72.5    4
```

〈데이터 출처 : <https://www.randomservices.org/random/data/Galton.txt>〉

4) 분석대상 데이터

● 데이터 속성

변수명	Family	Father	Mother	Gender	Height	Kids
설명	가족 번호 (ID)	아버지의 키 (인치)	어머니의 키 (인치)	성별 M : 남성 F : 여성	자녀의 키 (인치)	가족별 자녀들의 수

우리가 궁금한 것

아버지 키와 아들키의 상관성이
있을까?

〈데이터 출처 : <https://www.randomservices.org/random/data/Galton.txt>〉

4) 분석대상 데이터

```
hf <- read.table  
("https://www.randomservices.org/random/data/  
Galton.txt", header = T, stringsAsFactors = F)
```

▶ # 데이터 불러오기

```
stringsAsFactors = F)
```

▶ # 데이터 구조 확인

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

4) 분석대상 데이터

```
str(hf$Gender)
```

▶ # 성별 데이터 유형 확인 : character형

```
hf$Gender <- factor(hf$Gender, levels = c("M","F"))
```

▶ # factor 형의로 변형

```
str(hf$Gender)
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

4) 분석대상 데이터

```
hf.son <- subset(hf, Gender == "M")
```

▶ # 아들들만 선택

```
str(hf.son)
```

▶ # 새로운 데이터 프레임 구조 확인

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

4) 분석대상 데이터

```
par(mar = c(4,4,1,1))
```

▶ # 도화지 사이즈 설정

```
plot(hf.son$Father, hf.son$Height,  
     티뮤 = "아버지의 키", ylab = "아들의 키",  
     main = "아버지와 아들의 키", col = 3)
```

▶ # 아버지와 아들 키의 산점도

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

4) 분석대상 데이터

```
abline(v = mean(hf.son$Father), col = 2, lty = 2)
```

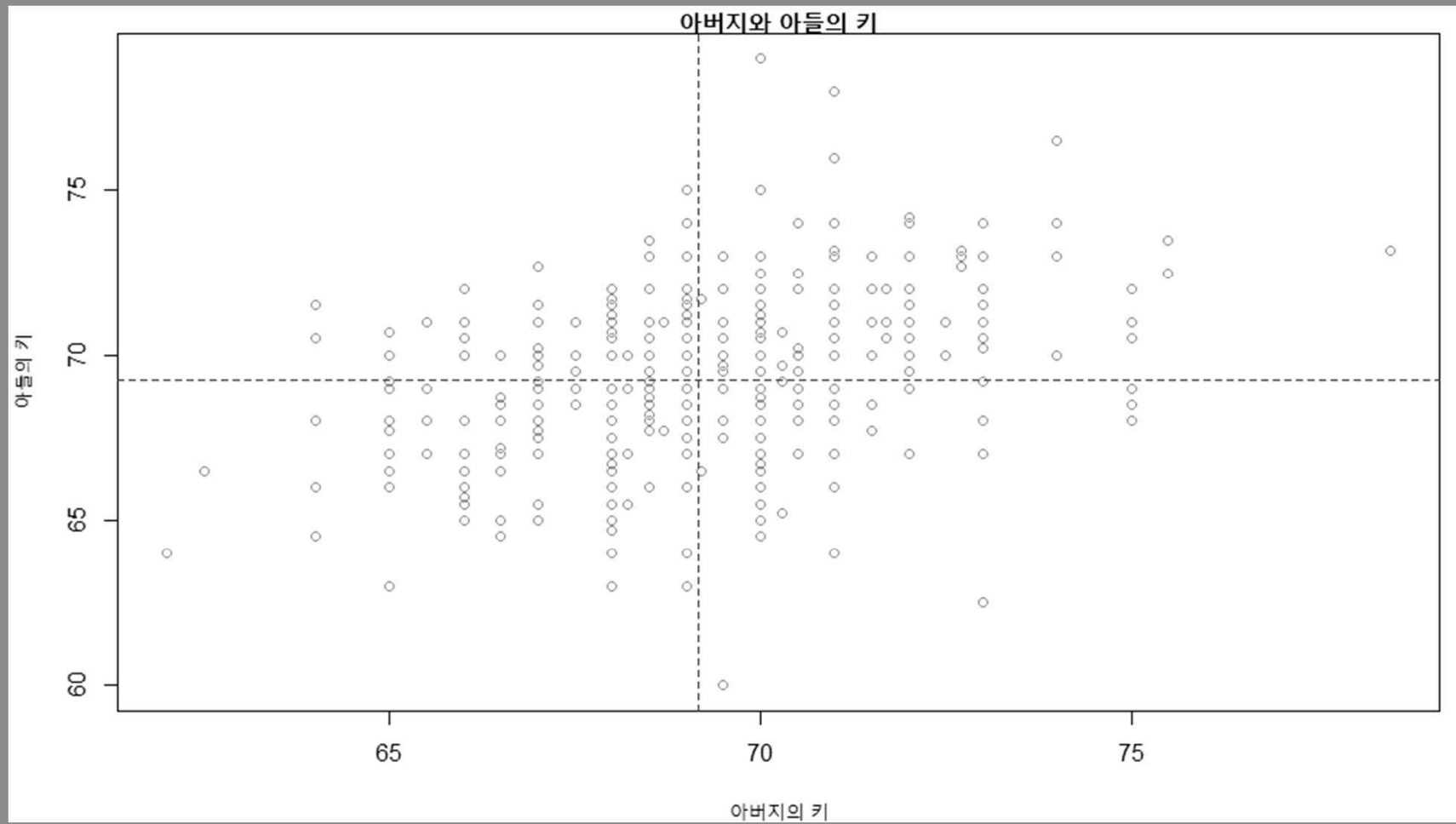
▶ # 아버지 키의 평균선

```
abline(h = mean(hf.son$Height), col = 2, lty = 2)
```

▶ # 아들 키의 평균선

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

▶ 산점도 결과



〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

6) 공분산식과 상관계수식

- 공분산식

$$Cov(X, Y)$$

$$= E[(X - E(X))(Y - E(Y))]$$

$$= E[(X - \mu_X)(Y - \mu_Y)],$$

$$E(X) = \mu_X,$$

$$E(Y) = \mu_Y$$

- 상관계수식

$$\rho_{XY}$$

$$= \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

$$= \frac{E[(X - E(X))(Y - E(Y))]}{\sigma_X \sigma_Y}$$

6) 공분산식과 상관계수식

```
1 f.mean <- mean(hf.son$Father)
```

▶ # 아버지들의 키 평균

```
2 s.mean <- mean(hf.son$Height)
```

▶ # 아들들의 키 평균

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

6) 공분산식과 상관계수식

3

```
cov.num <- sum( (hf.son$Father-f.mean) *  
(hf.son$Height - s.mean) )
```

▶ # 식을 활용한 공분산 구하기

4

```
(cov.xy <- cov.num / (nrow(hf.son) - 1))
```

▶ # 식을 활용한 공분산 구하기

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

6) 공분산식과 상관계수식

5 `cov(hf.son$Father, hf.son$Height)`

▶ # 함수를 활용한 공분산 구하기

6 `(r.xy <- cov.xy / (sd(hf.son$Father) *
sd(hf.son$Height)))`

▶ # 식을 활용한 상관계수 구하기

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

6) 공분산식과 상관계수식

```
7 cor(hf.son$Father, hf.son$Height)
```

▶ # 함수를 활용한 상관계수 구하기

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

7) 산점도 그리기

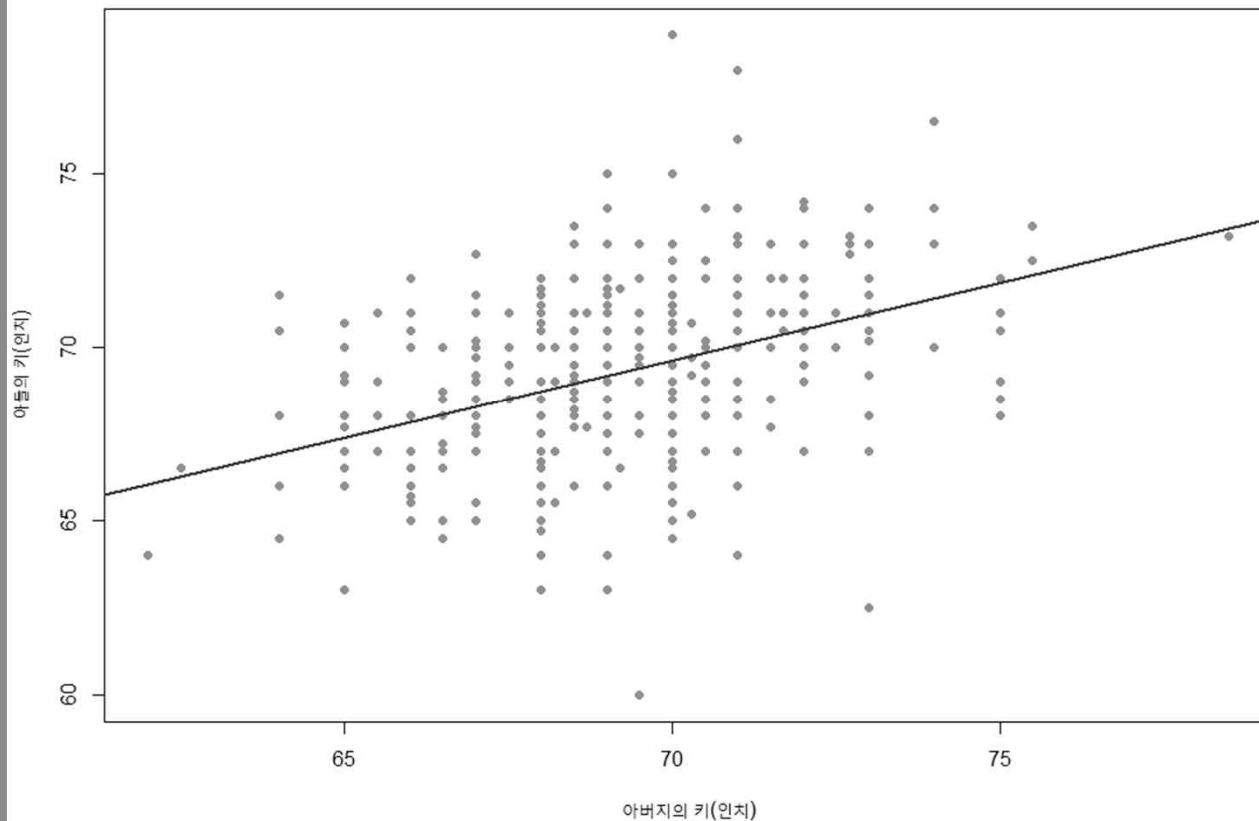
```
par(mfrow=c(1, 1), mar=c(4, 4, 1, 1))
```

```
plot(Height~Father,col = 3, pch=16, data=hf.son,  
     xlab="아버지의 키(인치)", ylab="아들의 키(인치)")
```

```
abline(lm(Height~Father, data=hf.son), col="red",  
       lwd=2)
```

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

▶ 산점도 그리기



“ 아버지와
아들의 키는
양의 선형관계를
가지고 있고
상관 계수는 0.39로
나타남 ”

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

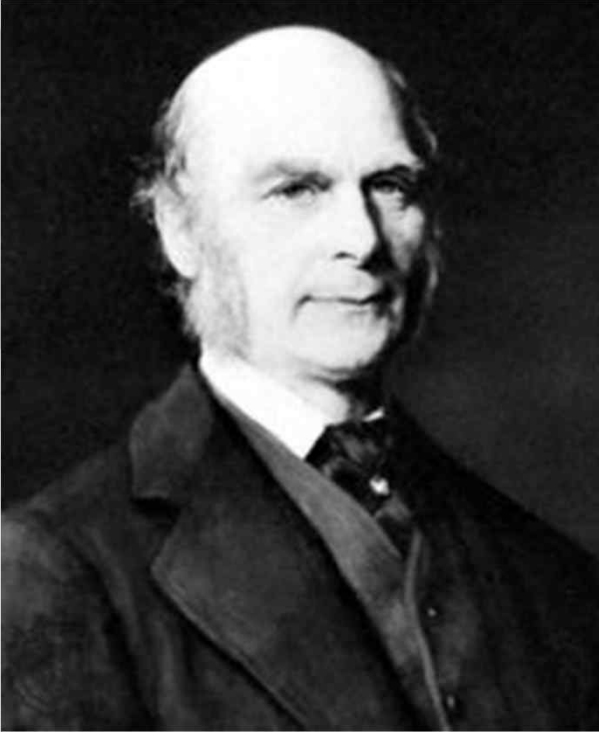


03

회귀분석

- | | |
|---------------|---------------|
| 1) 회귀분석이란 | 3) 단순선형 회귀 모형 |
| 2) 독립변수와 종속변수 | 4) 단순선형 회귀 분석 |

1) 회귀분석이란



평균으로의 회귀(Regression to The Mean) 현상을
증명하기 위해 회귀분석(Regression Analysis)을
만든 것으로 알려짐

- 생물학자 프랜시스 골턴(Francis Galton)

1) 회귀분석이란

평균으로의 회귀

부모와 아이의 키를 측정했을 때 극단적인 값이 나타나도,
그 다음에 새로 측정했을 때는 평균에
더 가까워지는(평균으로 회귀하는) 경향성을 보고
일반화한 용어

회귀분석 (Regression Analysis)

주어진 자료들이 어떤 특정한
경향성을 띠고 있다는
아이디어로부터 시작함

→ 기본적으로 변수들 사이에서 나타나는 경향성을
설명하는 것을 주 목적으로 함

2) 독립변수와 종속변수

● 인과관계

인과관계

원인과 결과 관계를 뜻하는 인과관계는 상관관계처럼 계산을 통해 구하는 것이 아닌, 주의 깊은 자료의 관찰을 통해 얻을 수 있는 관계

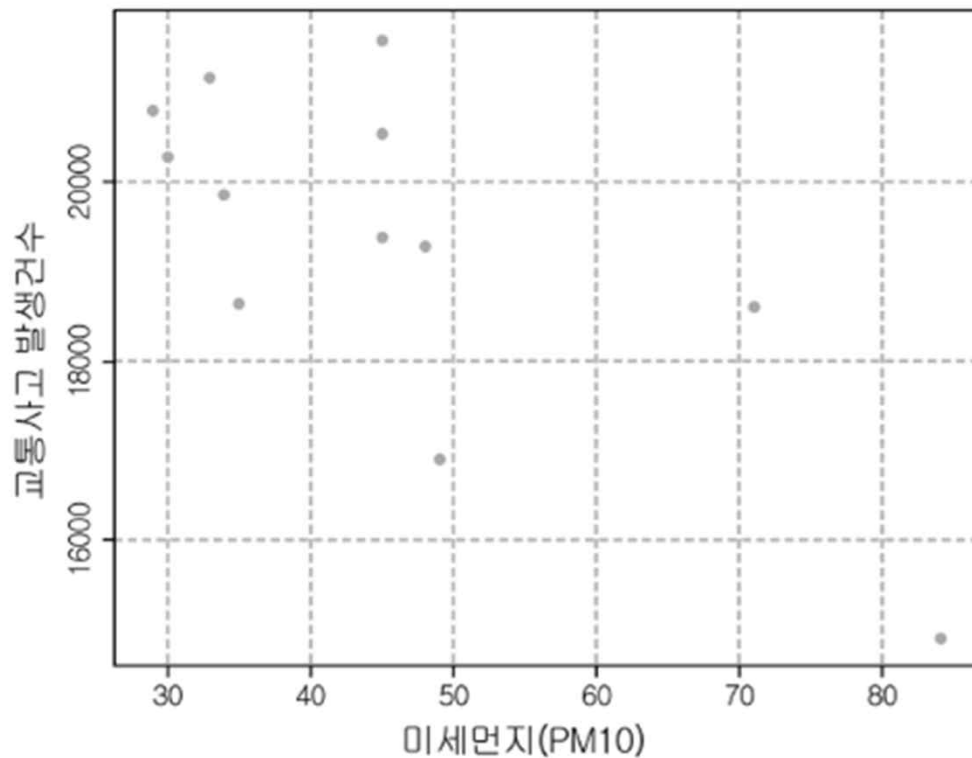


자료에 대한 깊은 통찰이 없다면 잘못된 인과관계를 도출할 수 있음

2) 독립변수와 종속변수

● 인과관계

미세먼지와 교통사고 데이터 산점도



- 미세먼지(PM-10) 농도에 따라 교통사고 발생이 어떤 연관이 있는지 알아보고자 작성한 도표
 - 전반적으로 미세먼지 농도가 짙어 질수록 교통사고 발생은 줄어드는 경향을 보임
- ➡ 이를 통해 미세먼지가 증가하면 교통사고 건수가 줄어든다고 할 수 있을까?

2) 독립변수와 종속변수

- 인과관계

- ▶ 관찰연구



사회현상은 관찰연구를 통해 연구하는 경우가 많음



이 경우에는 실험을 통제할 수 없음을 인정하고,
사전지식과 사회에 대한 깊은 통찰력을 가져야 함



2) 독립변수와 종속변수

- 인과관계
 - ▶ 관찰연구

“ 다음을 고민해 봅시다. ”

두 변수의
연관성

원인과
결과에
대한 고민

제3의
요인

2) 독립변수와 종속변수

● 인과관계

▶ 사례

아이스크림과
익사사고

어린이 시력과
전등불

국가 부채와
GDP

사과 수입과
이혼율

- 아이스크림 판매량이 증가할수록 익사사고 발생이 증가하였음
 - 즉, 익사사고 발생을 억제하기 위해 아이스크림의 판매를 금지해야 함
- ➡ 제3의 요인 : 계절

〈참조 : 위키피디아〉

2) 독립변수와 종속변수

● 인과관계

▶ 사례

아이스크림과
익사사고

어린이 시력과
전등불

국가 부채와
GDP

사과 수입과
이혼율

- 불을 켜고 자는 어린이의 경우,
나이가 들어 근시가 될 경우가 많음
 - 즉, 근시를 예방하기 위해 어릴 때부터
잠을 잘 때 불을 켜지 말아야 함
- ➡ 제3의 요인 : 부모의 근시

〈참조 : 위키피디아〉

2) 독립변수와 종속변수

● 인과관계

▶ 사례

아이스크림과
익사사고

어린이 시력과
전등불

국가 부채와
GDP

사과 수입과
이혼율

- 국가 부채가 GDP의 90% 이상이 될 경우 국가의 성장률이 느려짐
- 즉, 높은 국가 부채는 국가의 성장을 느리게 함

➡ 뒤바뀐 인과관계

〈참조 : 위키피디아〉

2) 독립변수와 종속변수

● 인과관계

▶ 사례

아이스크림과
익사사고

어린이 시력과
전등불

국가 부채와
GDP

사과 수입과
이혼율

- 사과의 수입이 증가할수록 이혼률이 증가함
 - 즉, 이혼률을 낮추기 위해 사과 수입을 금지해야 함
- ➡ 인과관계를 확인할 수 없는 두 변수

〈참조 : 위키피디아〉

2) 독립변수와 종속변수

- 인과관계

종속변수

변수 간의 관계에서
다른 변수에 의해 영향을
받아 그 값이 결정되는 변수

독립변수

변수 간의 관계에서
다른 변수에 의해 영향을
미치는 변수



3) 단순선형 회귀 모형

단순선형 회귀 모형

- 두 확률변수 X, Y 에서 X 가 독립변수이고, Y 가 종속변수일 경우 독립변수 X 의 개별값 x_1, x_2, \dots, x_n 에 대응하는 종속변수 Y 의 관찰값 y_1, y_2, \dots, y_n 에 대해 다음과 같은 모형

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$i = 1, 2, \dots, n, \epsilon_i \sim N(0, \sigma^2)$$

3) 단순선형 회귀 모형

- 회귀계수

회귀계수

앞의 식에서 두 상수 β_0, β_1 을 (모집단)회귀계수라고 함

↳ 각각 직선의 방정식에서 절편과 기울기의 역할을 함



두 상수는 미지의 모수로, 표본으로부터 추정을 통해 계산함

➡ 추정된 회귀계수를 이용하여 구한 식으로 나타나는 직선을 추정된 회귀직선이라고 함

4) 단순선형 회귀 분석

- 회귀계수의 추정

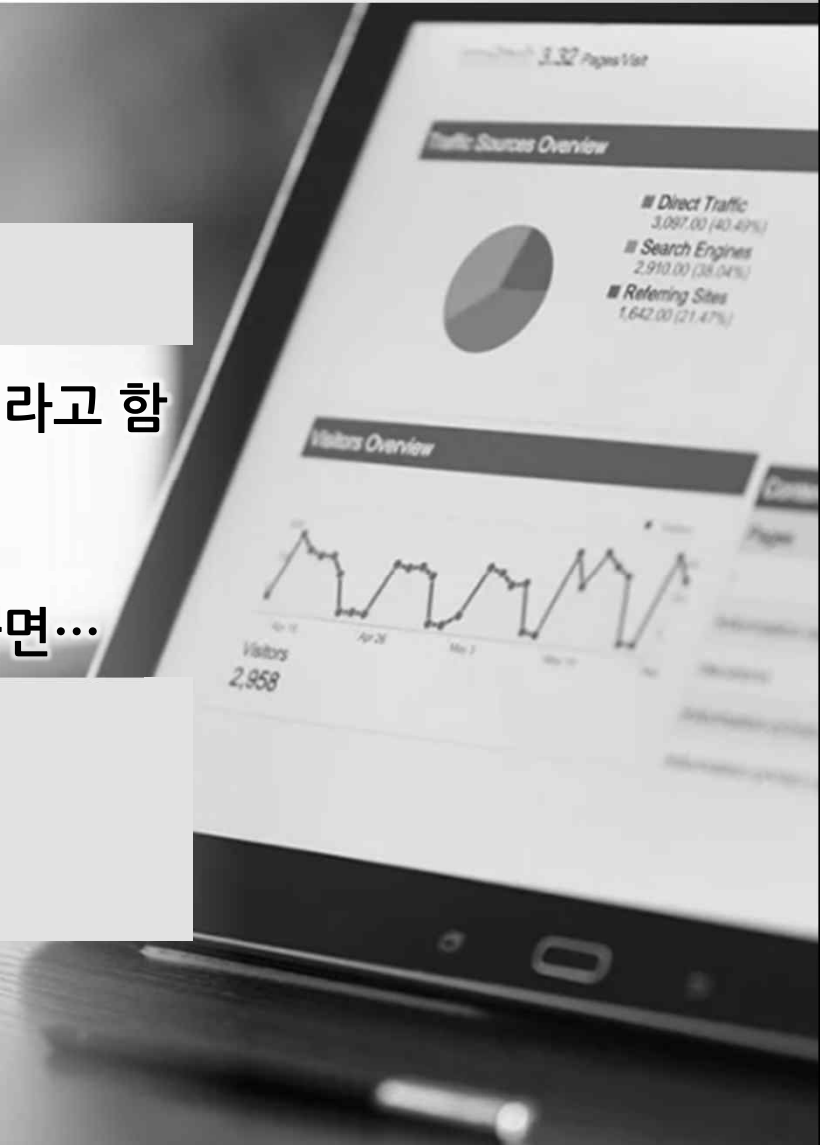
앞선 회귀식을 ϵ_i 에 대해 정리하면...

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

이 되고, ϵ_i 를 오차 혹은 오차항이라고 함

회귀계수의 추정은 오차들의 제곱합을 이용하여 계산하면...

$$S = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$



4) 단순선형 회귀 분석

- 회귀계수의 추정

- ▶ 최소제곱법과 최소제곱추정량

최소제곱추정량

회귀계수추정의 한 방법으로 오차들의 제곱합을
최소로 하는 β_0, β_1 의 추정량인 b_0, b_1 를 구하는
최소제곱법을 통해 구한 추정량



4) 단순선형 회귀 분석

● 추정된 회귀직선



회귀계수에 대한 추정량 b_0, b_1 과 종속변수 Y 의 예측 값을 \hat{y} 이라 하면, 추정된 회귀직선은 다음과 같음



$$\hat{y} = b_0 + b_1x$$

또한 $b_0 = \bar{y} - b_1\bar{x}$ 이므로...

$$\begin{aligned}\hat{y} &= b_0 + b_1x = \bar{y} - b_1\bar{x} + b_1x \\ &= \bar{y} + b_1(x - \bar{x})\end{aligned}$$



추정된 회귀직선을 통해 독립변수가 가질 수 있는 값에 대응하는 종속변수의 값을 추측할 수 있음

4) 단순선형 회귀 분석

```
1 mean.x <- mean(hf.son$Father)
```

▶ # 아버지들의 키 평균

```
2 mean.y <- mean(hf.son$Height)
```

▶ # 아들들의 키 평균

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

4) 단순선형 회귀 분석

3

```
sxy <- sum((hf.son$Father - mean.x) *  
(hf.son$Height - mean.y))
```

▶ # 아버지의 키의 편차와 아들키의 편차들의 곱의 합

4

```
sxx <- sum((hf.son$Father - mean.x)^2)
```

▶ # 아버지의 키의 편차 제곱합

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

4) 단순선형 회귀 분석

```
5 ( b1 <- sxy / sxx )
```

▶ # 회귀계수 추정치(기울기)

```
6 ( b0 <- mean.y - b1 * mean.x )
```

▶ # 회귀 계수 추정치(절편)

```
7 lm(Height ~ Father, data = hf.son)
```

▶ #회귀 함수를 활용하여 회귀식 구하기

〈소스코드 : https://github.com/LEESUAJE1978/r_statistics/blob/master/lecture4.R〉

정리 하기

가설검정(Hypothesis Test)

✓ 가설검정 단계

- 1단계 : 가설 수립
- 2단계 : 표본으로부터 검정을 위한 통계량 계산
- 3단계 : 가설 선택의 기준 수립
- 4단계 : 판정

✓ 영가설 (Null)과 대립가설 (Alternative)

- 영가설 (귀무가설, H_0)
: 기존에 알려진 것과 차이가 없음을 나타냄
- 대안가설 (대립가설, H_1)
: 연구자가 밝히고자 하는 가설로
연구 가설이라고도 함

정리 하기

통계분석(Statistics Analysis)

- ✓ 상관분석(Correlation Analysis)
 - 두 변수 간에 어떤 선형적 또는 비선형적 관계를 가지고 있는지 분석하는 방법
 - 두 변수는 서로 독립적인 관계이거나 상관된 관계일 수 있으며, 이 때 두 변수 간의 관계의 강도를 상관관계(Correlation Coefficient)라고 함
 - 상관관계의 정도를 나타내는 단위로 모상관계수로 ρ ($\rho\omega$ 로), 표본 상관 계수로 r 을 사용함

정리 하기

통계분석(Statistics Analysis)

- ✓ 회귀분석(Regression Analysis)
 - 회귀분석(Regression Analysis)는 생물학자 프랜시스 골턴(Francis Galton)이 평균으로의 회귀(Regression to The Mean)현상을 증명하기 위해 만든 것으로 알려짐
 - 기본적으로 변수들 사이에서 나타나는 경향성을 설명하는 것을 주 목적으로 함

The background is a dark, abstract composition featuring a network of glowing white lines and geometric shapes. A prominent hexagon is located in the upper right quadrant, with lines radiating from it. On the left side, there are several overlapping, parallel lines that form a series of nested, elongated shapes. The overall aesthetic is futuristic and technological.

- 다음 시간에 살펴 볼 내용 -

11강 빅데이터 분석 과제 도출

수고하셨습니다.