

고고학 자료 통계분석

Week 3 : 기초 통계(2)

숭실대학교 사학과 석사과정 1학기
주 찬 혁

계획

주차	제목	내용
1	Intro	소개, R 설치
2	기초 통계(1)	모집단과 표본, 기술통계량
3	기초 통계(2)	변수의 종류, 가설과 검정, 오류, 분석절차
4	전처리	데이터 전처리
5	시각화	다양한 종류의 그래프
6	검정	t-검정, Chi-square 검정, Median Polish
7	회귀분석	선형회귀, 다중선형회귀, 로지스틱회귀
8	군집분석	K-means,
9	판별분석	DA, MDA
10	주성분분석	PCA

복습

- 모집단과 표본이 무엇인지 안다.
- **고고학적 맥락을 살피면서 통계분석을 해야 한다.**
- 기술통계량이 무엇인지 안다.
- R로 기술통계량을 산출할 수 있다.

변수의 종류

- 수량적 특성에 따른 분류
 - 질적 변수, 양적 변수
- 자료의 특성에 따른 분류
 - 범주형 변수, 연속형 변수
- 자료의 상관관계에 따른 분류
 - 독립변수, 종속변수

***변수의 종류에 따라 데이터 수집, 분석 방법 등이 설정됨**

수량적 특성에 따른 분류

- 질적 변수, 양적 변수

질적 변수 (Qualitative Variable)	서열 질적 변수 (Ordered-Qualitative Variable)
	비서열 질적 변수 (Unordered-Qualitative Variable)
양적 변수 (Quantitative Variable)	연속형 변수 (Continuous Variable)
	비연속형 변수 (UnContinuous Variable)

질적 변수

서열 질적 변수

조사요원별 자격기준
조사단장
책임조사원
조사원
준조사원
보조원

6세기 이후 신라 묘제
횡혈식석실분
석관묘, 석곽묘

학위
박사
석사
학사

신라 골품제
성골
진골
6두품
...

구성 요소 사이에
위계(서열)가 존재

비서열 질적 변수

성별	
남성	여성

삼국		
고구려	백제	신라

반려동물 종류					
개	고양이	앵무새	뱀	물고기	거북이

석기 종류			
주먹도끼	긁개	찌르개	...

모든 구성요소가 **평등**

양적 변수

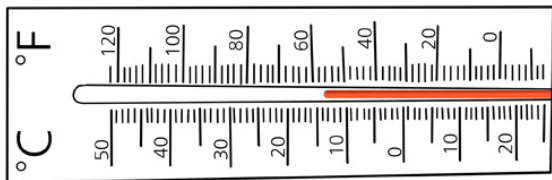
연속형 변수

키
190cm ~
180~189cm
170~179cm
160~169cm
150~159cm
...

나이
60대 ~
50대
40대
30대
20대
10대

수로 표현되고
구성요소가 연속됨
*순서는 위계가 아님

온도



비연속형 변수

손가락 개수
10개

아파트 층수
15층

A유적의 유물 출토량					
토기	석기	청동기	철기	불상	인골
10개	5개	1개	1개	2기	1구

정수로 표현되고
셀 수 있음

자료의 특성에 따른 분류

- 범주형 변수, 연속형 변수

범주형 변수=이산형 변수 (Categorical Variable = Discrete Variable)	명목 변수 (Nominal Variable)
	순위 변수 (Ordinal Variable)
연속형 변수 (Continuous Variable)	간격 변수 (Interval Variable)
	비율 변수 (Ratio Variable)

범주형 변수

질적변수, 비연속형 변수 포함 / 모든 구성요소는 구분됨

명목형 변수

토기 돌대 유무	
있음	없음

주거지 형태		
장방형	방형	원형

형식분류			
1	2	3	4

*숫자는 크기의 성질이 없음

A유적 출토 유물					
토기	석기	청동기	철기	불상	인골

묘제					
석실묘	석곽묘	석관묘	목곽묘	목관묘	...

구성요소가 범주에 속해지고 이름이 부여됨

순위 변수

성적
A
B
C
D
F

학위
박사
석사
학사

6세기 이후 신라 묘제
횡혈식석실분
석관묘, 석곽묘

명목형 변수에 위계(순위)를 부여하여
순위 변수로 사용할 수 있음(반대도 O)

연속형 변수

비율 변수

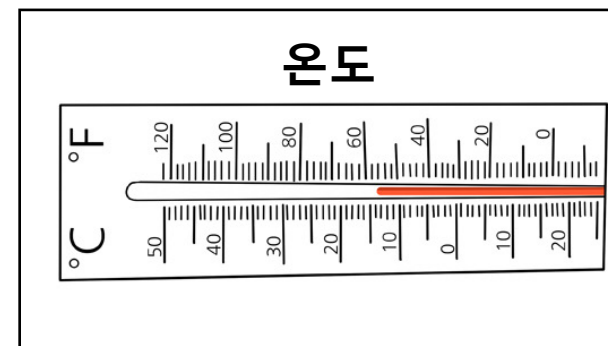
키
190cm ~
180~189cm
170~179cm
160~169cm
150~159cm
...

나이
60대 ~
50대
40대
30대
20대
10대

0과의 비율이 의미 **있음**

Tip : 0은 절대적

간격 변수



0과의 비율이 의미 **없음**

Tip : 0은 인간이 임의적으로 부여한 존재

자료의 상관관계에 따른 분류

- 독립변수, 종속변수

독립변수(Independent Variable)
종속변수(Dependent Variable)

자료의 상관관계에 따른 분류

$$y = a + bx$$

x -> 독립변수 / y -> 종속변수

x에 b만큼 변화가 가해지면
y는 그 영향으로 인해 변함

Feature Number	Excavation institute	BP	Error	From	To	Median	Type
Proto-three Kingdoms~Three Kingdoms 4	Hanbaek Cultural Property Research Institute	1730	40	220	405	309	마한
마-A-26	Hanbaek Cultural Property Research Institute	1590	30	406	542	480	마한
마-A-28	Hanbaek Cultural Property Research Institute	1650	30	264	533	398	마한
마-A-36	Hanbaek Cultural Property Research Institute	1630	30	346	536	418	마한
마-A-63	Hanbaek Cultural Property Research Institute	1700	30	253	406	348	마한
마-A Pottery	Hanbaek Cultural Property Research Institute	1860	40	64	243	156	마한

독립변수 : 다른 변수에 영향을 주는 변수
- 연구자가 조절하는 대상

종속변수 : 다른 변수에게 영향을 받는 변수
- 연구자가 알고 싶은 대상

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

R의 자료형

숫자형(Numeric)	논리형(Logical)
문자형(Character)	팩터형(Factor)

R의 자료형

as.numeric(변수)

as.logical(변수)

as.character(변수명)

as.factor(변수명)

> as.□□□의 자료형으로 변환

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for variable assignment and conversion.

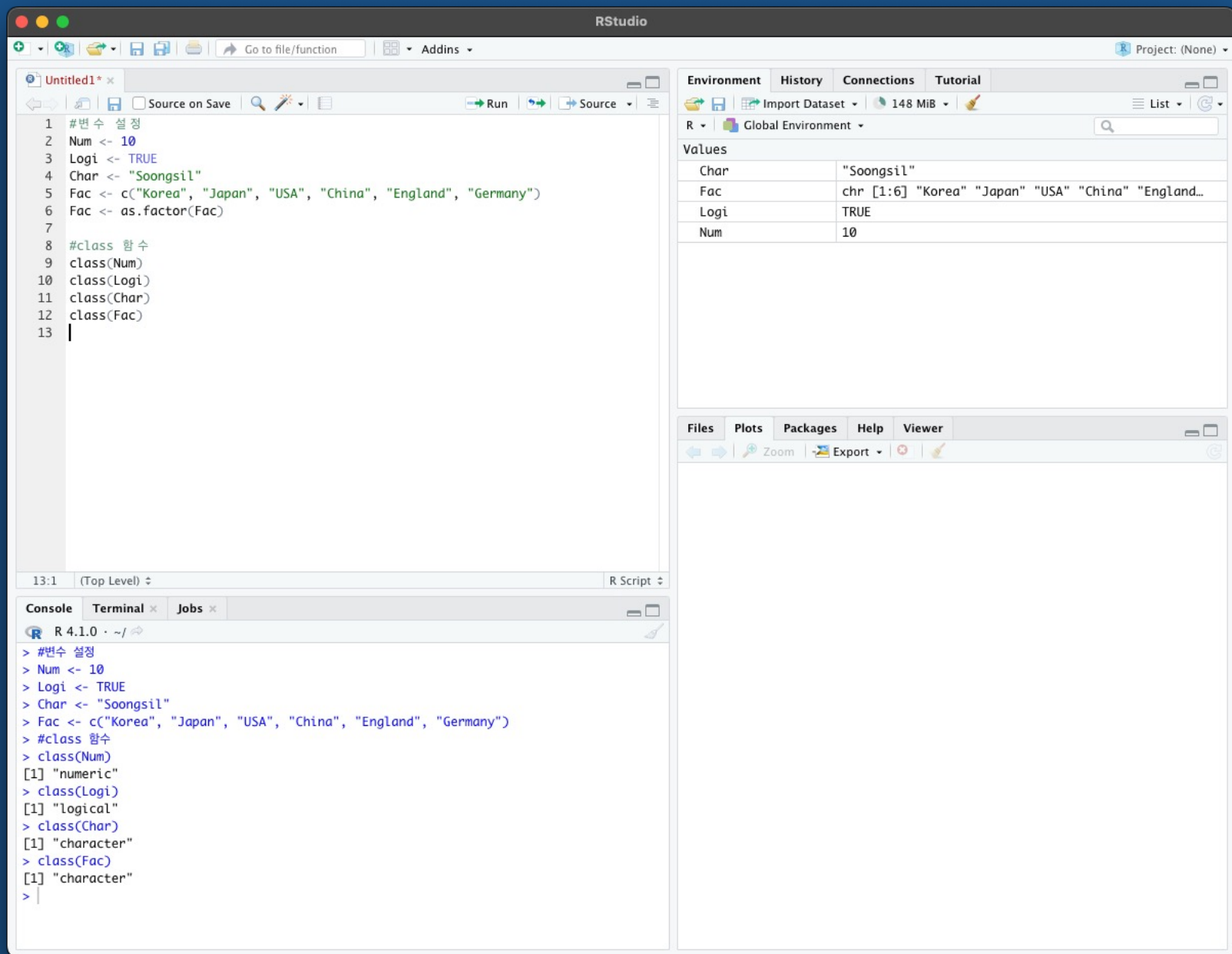
```
1 #변수 설정
2 Num <- 10
3 Logi <- TRUE
4 Char <- "Soongsil"
5 Fac <- c("Korea", "Japan", "USA", "China", "England", "Germany")
6 Fac <- as.factor(Fac)
```
- Environment Pane:** Displays the current environment with the following values:

Variable	Value
Char	"Soongsil"
Fac	chr [1:6] "Korea" "Japan" "USA" "China" "England..."
Logi	TRUE
Num	10
- Console:** Shows the execution of the code from the source editor.

```
> #변수 설정
> Num <- 10
> Logi <- TRUE
> Char <- "Soongsil"
> Fac <- c("Korea", "Japan", "USA", "China", "England", "Germany")
>
```

R의 자료형

*class
: 자료형 출력



The screenshot displays the RStudio interface with the following components:

- Source Editor (Untitled1.R):** Contains R code for variable assignment and class checking.

```
1 #변수 설정
2 Num <- 10
3 Logi <- TRUE
4 Char <- "Soongsil"
5 Fac <- c("Korea", "Japan", "USA", "China", "England", "Germany")
6 Fac <- as.factor(Fac)
7
8 #class 함수
9 class(Num)
10 class(Logi)
11 class(Char)
12 class(Fac)
13 |
```
- Environment Pane:** Shows the Global Environment with the following values:

Variable	Value
Char	"Soongsil"
Fac	chr [1:6] "Korea" "Japan" "USA" "China" "England..
Logi	TRUE
Num	10
- Console:** Shows the output of the R code execution.

```
> #변수 설정
> Num <- 10
> Logi <- TRUE
> Char <- "Soongsil"
> Fac <- c("Korea", "Japan", "USA", "China", "England", "Germany")
> #class 함수
> class(Num)
[1] "numeric"
> class(Logi)
[1] "logical"
> class(Char)
[1] "character"
> class(Fac)
[1] "character"
> |
```


R의 자료형

*str
: 자료형과 구성요소 출력

The screenshot displays the RStudio interface with the following components:

- Source Editor (Untitled1.R):** Contains R code for variable assignment and class checking.
- Environment Pane:** Shows the current environment with variables Char, Fac, Logi, and Num.
- Console:** Displays the output of the R commands.

R Code (Source Editor):

```
1 #변수 설정
2 Num <- 10
3 Logi <- TRUE
4 Char <- "Soongsil"
5 Fac <- c("Korea", "Japan", "USA", "China", "England", "Germany")
6 Fac <- as.factor(Fac)
7
8 #class 함수
9 class(Num)
10 class(Logi)
11 class(Char)
12 class(Fac)
13
14 #str 함수
15 str(Num)
16 str(Logi)
17 str(Char)
18 str(Fac)
19
```

Environment Pane (Values):

Variable	Value
Char	"Soongsil"
Fac	chr [1:6] "Korea" "Japan" "USA" "China" "England..
Logi	TRUE
Num	10

Console Output:

```
[1] "numeric"
> class(Logi)
[1] "logical"
> class(Char)
[1] "character"
> class(Fac)
[1] "character"
> #str 함수
> str(Num)
num 10
> str(Logi)
logi TRUE
> str(Char)
chr "Soongsil"
> str(Fac)
chr [1:6] "Korea" "Japan" "USA" "China" "England" "Germany"
>
```

가설과 검정 그리고 오류

- 가설

- 귀무가설(영가설, H_0)
- 대립가설(대안가설, H_1)

- 검정

- 채택(Accept)할 것인가?, 기각(Reject)할 것인가?

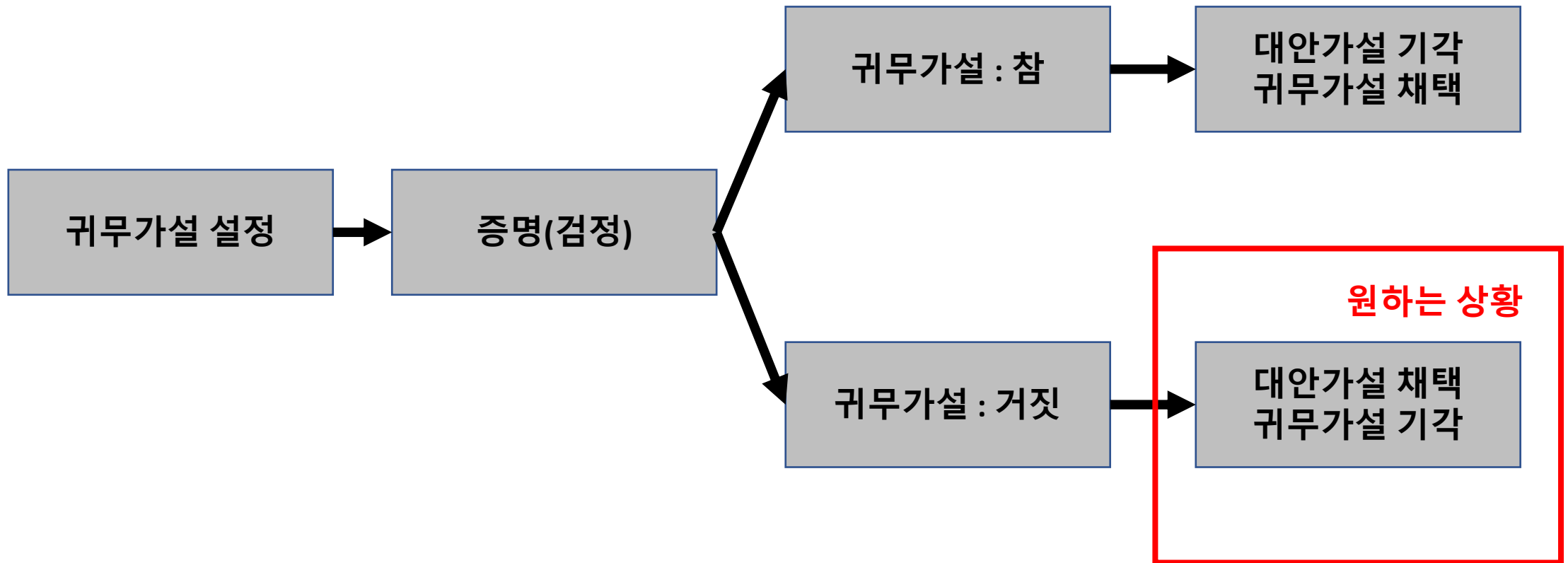
- 오류

- 1종 오류
- 2종 오류

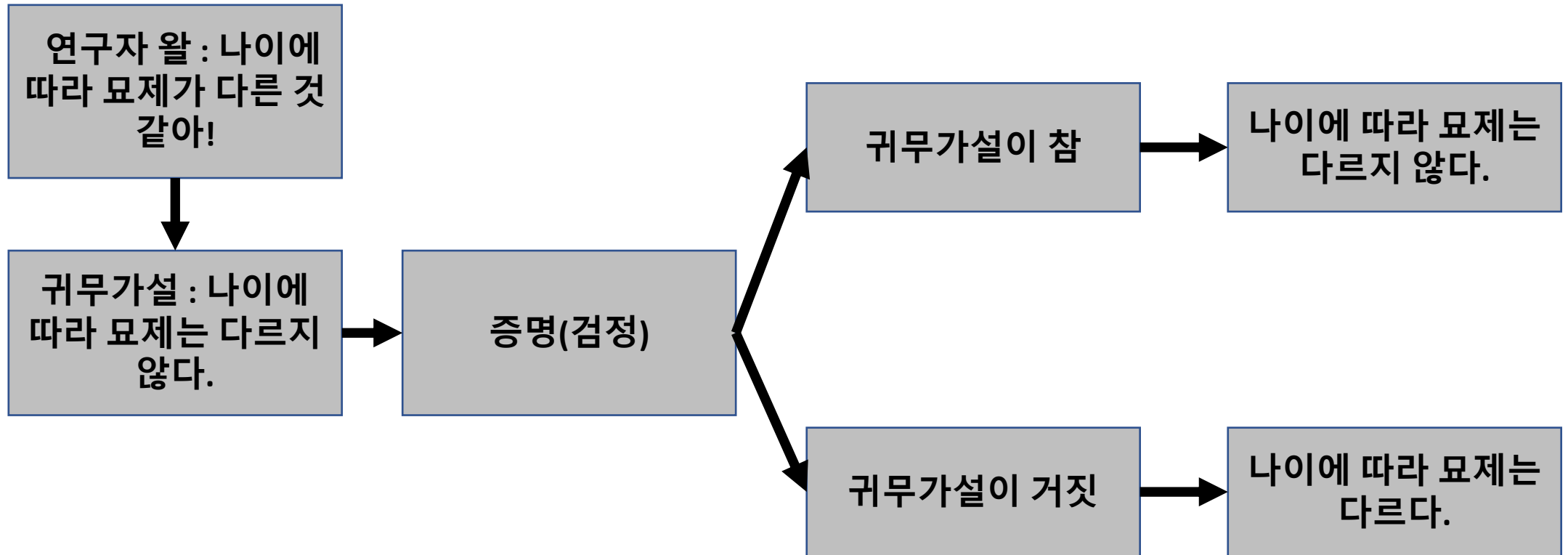
가설과 검정

- 모든 것은 참 아니면 거짓이라는 전제
- 한 명제에 대한 가설을 세우고, 이 가설이 거짓임을 증명하여 명제가 참임을 증명(검정) -> 다양한 검정 방법은 6주차에...
- 여기서 처음 세워지는 가설이 귀무가설(영가설), 이에 반대되는 가설이 대립가설(대안가설)

가설과 검정



가설과 검정



오류

- 어느 가설 검정도 100% 확실하지 않음에 따라 오류가 발생
- 1종 오류 : 귀무가설이 참인데 기각
- 2종 오류 : 귀무가설이 거짓인데 채택

오류

***1종 오류와 2종 오류의 가능성은 항상 존재함
어떤 오류를 줄일지는 상황에 따라 판단**

연구 주제 : 나이에 따라 묘제가 다르다.

- 귀무가설 : 나이에 따라 묘제가 다르지 않다.
- 대립가설 : 나이에 따라 묘제가 다르다.

1종 오류

- 실제 : 나이에 따라 묘제가 다르지 않음
- 검정결과 : 나이에 따라 묘제가 다름

-> 없는 현상을 있다고 하는 것

2종 오류

- 실제 : 나이에 따라 묘제가 다름
- 검정결과 : 나이에 따라 묘제가 다르지 않음

-> 있는 현상을 없다고 하는 것

오류

출처 : <https://support.minitab.com/>

한 의료 분야 연구자가 두 약품의 효과를 비교하려고 합니다. 귀무 가설과 대립 가설은 다음과 같습니다.

- 귀무 가설(H_0): $\mu_1 = \mu_2$

두 약품의 효과가 동일합니다.

- 대립 가설(H_1): $\mu_1 \neq \mu_2$

두 약품의 효과가 동일하지 않습니다.

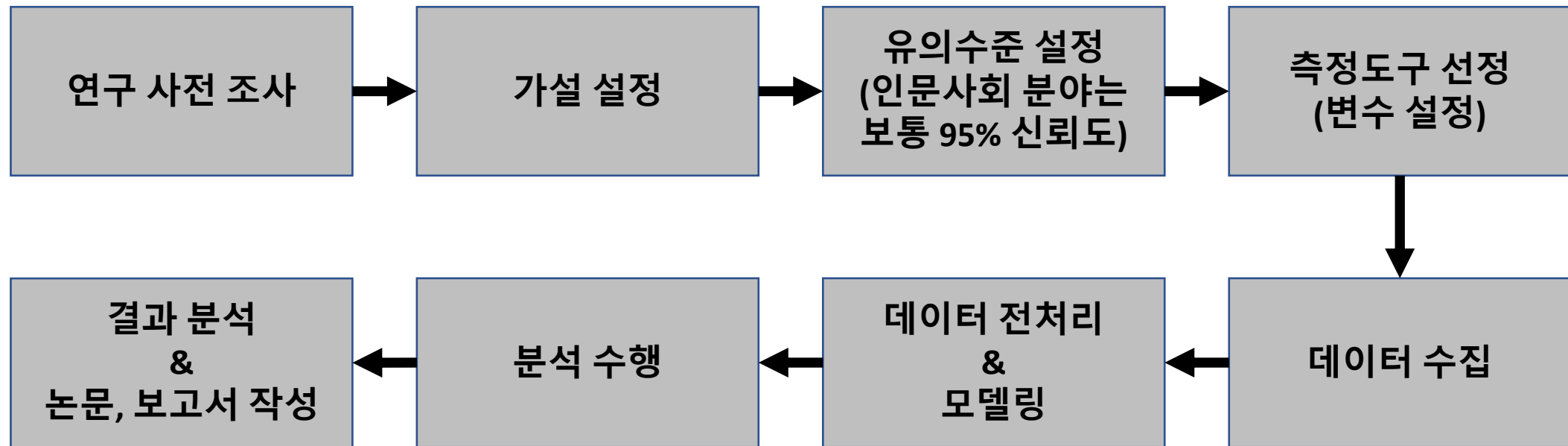
두 약품이 다르지 않지만 연구자가 귀무 가설을 기각하고 두 약품이 다르다는 결론을 내리는 경우 제1종 오류가 발생합니다. 약품의 효과가 동일한 경우, 환자는 어느 약품을 복용하거나 관계 없이 동일한 수준의 효과를 얻기 때문에 연구자는 이 오류가 매우 심각한 것으로 생각하지 않을 수도 있습니다. 그러나 제2종 오류가 발생하면 연구자는 귀무 가설을 기각해야 하지만 기각하지 못합니다. 즉, 연구자는 두 약품이 실제로는 다르지만 같다는 결론을 내리는 것입니다. 이 오류는 효과가 더 낮은 약품이 효과가 더 높은 약품 대신 판매될 경우 잠재적으로 생명을 위협할 수도 있습니다.

***각 오류의 위험성을 인지하고
상황에 맞게 적절히 조정해야함**

분석 절차

- **확증적 데이터 분석**
 - 추론통계 : 검정, 신뢰구간 추정 등...
- **탐색적 데이터 분석**
 - 기술통계 : 기술통계량, 데이터 분포 등...

확증적 데이터 분석(Confirmatory Data Analysis, CDA)



탐색적 데이터 분석(Exploratory Data Analysis, CDA)

