2021학년도 1학기 의학통계학 실습강의

경상국립대학교 의과대학 예방의학교실
김미지

주차	일시	강의내용	강의교수	비고
7	4. 14. (수) 09:00-11:50	8. 데이터 수집 및 정리	김미지	비대면 강의(실시간)
8	4. 21. (수) 09:00-11:50	중간고사	김미지	대면 시험
9	4. 28. (수) 09:00-11:50	9. 두 집단 간 차이 비교: t 검정, 카이제곱 검정, 비모수 검정	김미지	비대면 강의(실시간)
10	**5. 05. (수)** **09:00-11:50**	10. 세 집단 간 차이 비교: ANOVA, 사후분석, 비모수 검정	김미지	비대면 강의(동영상)
11	5. 12. (수) 09:00-11:50	11. 단순 연관성 확인: 상관분석	김미지	비대면 강의(실시간)
12	**5. 19. (수)** **09:00-11:50**	12. 인과적 연관성 확인: 회귀분석	김미지	비대면 강의(동영상)
13	5. 26. (수) 09:00-11:50	13. 빅데이터 분석 및 생물정보학 (bioinformatics) Ⅰ	유원기	비대면 강의(실시간/동영상) **과제 제출**
14	6. 02. (수) 09:00-11:50	14. 빅데이터 분석 및 생물정보학 (bioinformatics) Ⅱ	유원기	비대면 강의(실시간/동영상)
15	6. 09. (수) 09:00-11:50	기말고사	김미지	대면 시험(예정)

8. 데이터 수집 및 정리

2021-04-14 (수)

1) 통계 프로그램의 선택

다양한 통계 프로그램

Excel

SPSS

SAS

STATA

R

Python

JASP

Kolabtree

Top 10 Statistical Tools Used in Medical Research

S/N	Product	Developer	Learning Curve	Cost (USD)	Open Source	Software license	Interface	Written in	Most Common Use Cases
1	STATA	StataCorp LLC	Steep	Academic starting at $595/ industry starting at $1,195	No	Proprietary	CLI/GUI	C	Clinical Data Analysis & Public Health
2	R	R Foundation	Steep	Free	Yes	GNU/PL	CLI/GUI	C with chunks in Fortran/C++	Meta-Analysis using special packages (Metafor & JASP)
3	GraphPad Prism	GraphPad Software, Inc.	Shallow	595	No	Proprietary	GUI	C/C++	Biological Labs, Research & Clinical Data Analysis
4	SAS	SAS Institute	Pretty steep	~$6000 per seat (PC version)/~$28K per processor (Windows server) first-year fees for BASE, STAT, GRAPH, and ACCESS modules. Modules are licensed individually. Subsequent year fees are roughly half.	No	Proprietary	CLI/GUI	C	Clinical Data Analysis, Health & Life Sciences
5	IBM SPSS	IBM	Shallow	$4,975	No	Proprietary	CLI/GUI	Java	Systematic Reviews, Surveys & Clinical Data Analysis
6	MATLAB	MathWorks	Pretty steep	$2150 (commercial), $99 (student), toolboxes additional	No	Proprietary	CLI	C++ & Java	Meta-Analysis & Clinical Data Analysis
7	JMP	SAS Institute	Shallow	$1995 (commercial) $29.95/$49.95 (student) $495 for H.S. site licence	No	Proprietary	CLI/GUI	C++	Clinical Data Visualisation and Analysis
8	Minitab	Minitab Inc.	Shallow	$895-$1395 perpetual, $542 or less concurrent annual, $29.99/$49.99/$99.99 academic	No	Proprietary	CLI/GUI	Fortran	Clinical Data Analysis & Healthcare Analytics
9	STATISTICA	StatSoft	Steep	>$695	No	Proprietary	GUI	C	Clinical Data Visualisation and Analysis
10	Excel	Microsoft Corporation	Shallow	$8.25 per month	Yes	Proprietary	GUI	C and C++ and C#	Clinical Data Analysis & Meta-Analysis (MetaXL add-in)

Note: ***Pretty Steep*** = Very difficult and gradual learning curve, ***Steep*** = Difficult and gradual learning curve, ***Shallow*** = Relatively easy and quick to learn

R과 RStudio

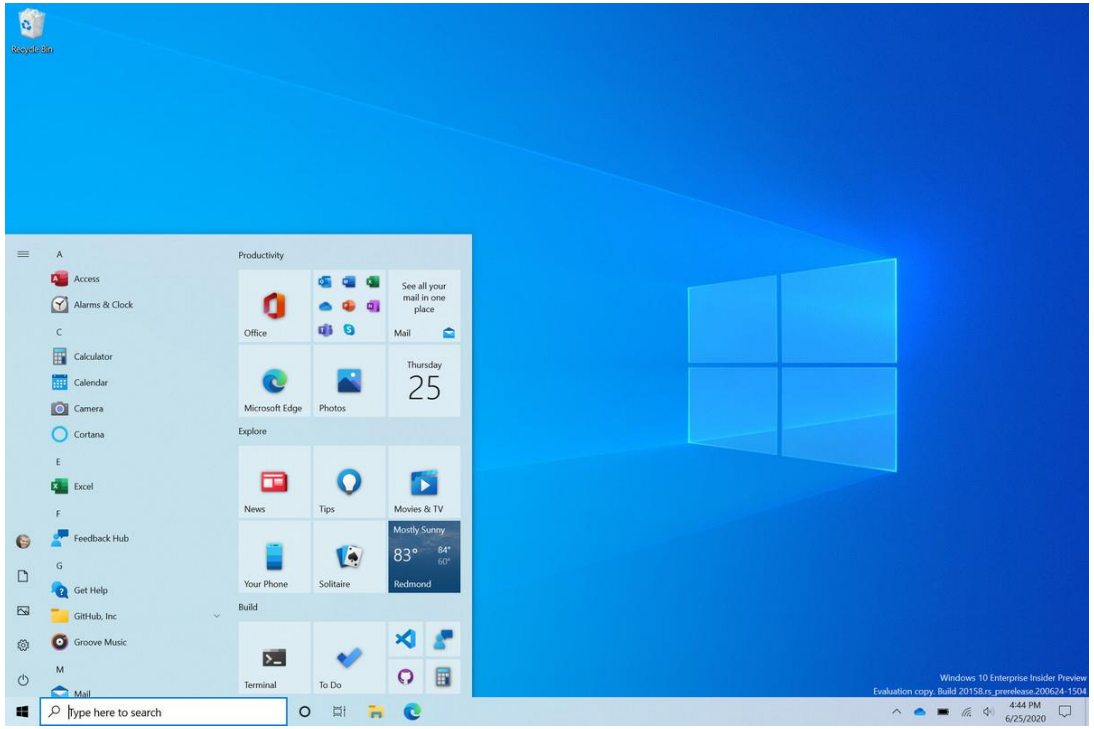
- R: 통계분석을 위한 통계 프로그램(무료)
- RStudio: R을 더욱 편리하게 사용하기 위한 인터페이스를 제공하는 프로그램(무료)
- 인터넷 연결이 필요함

R과 RStudio 설치

- 내 컴퓨터의 운영체제 확인
- R: https://cran.r-project.org/
- RStudio: https://rstudio.com/products/rstudio/download/
- 위의 홈페이지에서 무료 다운로드 가능

명령어(command) 기반 인터페이스

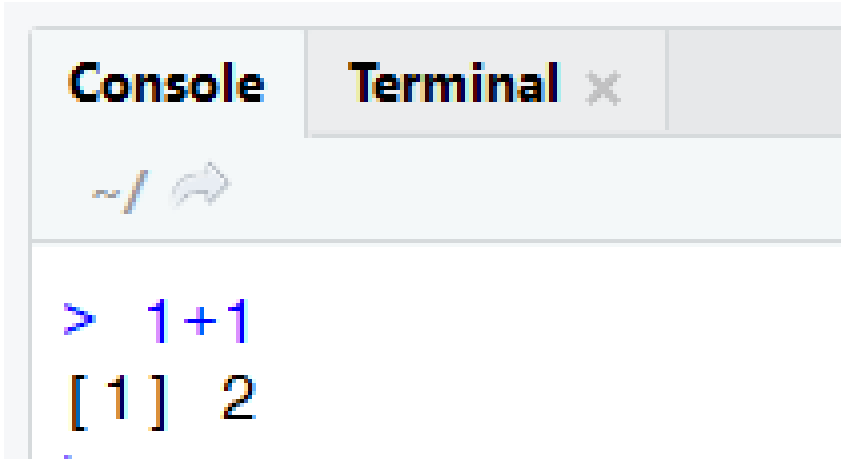
Windows



DOS

```
Welcome to FreeDOS


CuteMouse v1.9.1 alpha 1 [FreeDOS]
Installed at PS/2 port
C:\>ver

FreeCom version 0.82 pl 3 XMS_Swap [Dec 10 2003 06:49:21]

C:\>dir
 Volume in drive C is FREEDOS_C95
 Volume Serial Number is 0E4F-19EB
 Directory of C:\

FDOS                <DIR>  08-26-04  6:23p
AUTOEXEC BAT          435  08-26-04  6:24p
BOOTSECT BIN          512  08-26-04  6:23p
COMMAND  COM       93,963  08-26-04  6:24p
CONFIG   SYS          801  08-26-04  6:24p
FDOSBOOT BIN          512  08-26-04  6:24p
KERNEL   SYS       45,815  04-17-04  9:19p
        6 file(s)        142,038 bytes
        1 dir(s)   1,064,517,632 bytes free

C:\>_
```

명령어(command) 기반 인터페이스



R basic rules

- 대소문자를 구별함
- 변수명에 “-”이나 “ (space)”를 포함할 수 없음
- 큰 따옴표(“ ”) 안의 내용은 문자로 인식함
- 패키지를 한 번 설치하면 다시 설치할 필요 없이 “library” 또는 “require” 함수로 불러오면 됨
- 함수에 대해 궁금하다면 “?함수이름“ 또는 “help(함수이름)”으로 검색이 가능함
- “=“ 또는 “<-”으로 데이터를 입력(할당)할 수 있음
- 경로에 있는 “/”는 “//”이나 “＼”로 바꿔야 함
- “#” 뒤에 오는 것은 모두 실행하지 않음
- 명령어는 한 줄에 한 개씩 써야 함, 같은 줄 안에서 구분하고 싶다면 “;”로 구분함
- ()는 함수에 사용, []은 행렬의 위치를 지정하거나 조건문을 쓸 때 사용함

R data structure

- 데이터 이름: df
- 데이터 차원: 5*4(개체 5명, 변수 4개)
- 변수명: id, year, sex, age

id	year	sex	age
A901020501	2018	1	80
A901021301	2018	1	78
A901021302	2018	2	75
A901022601	2018	2	43
A901023303	2018	2	80

(특정 변수 선택하기)

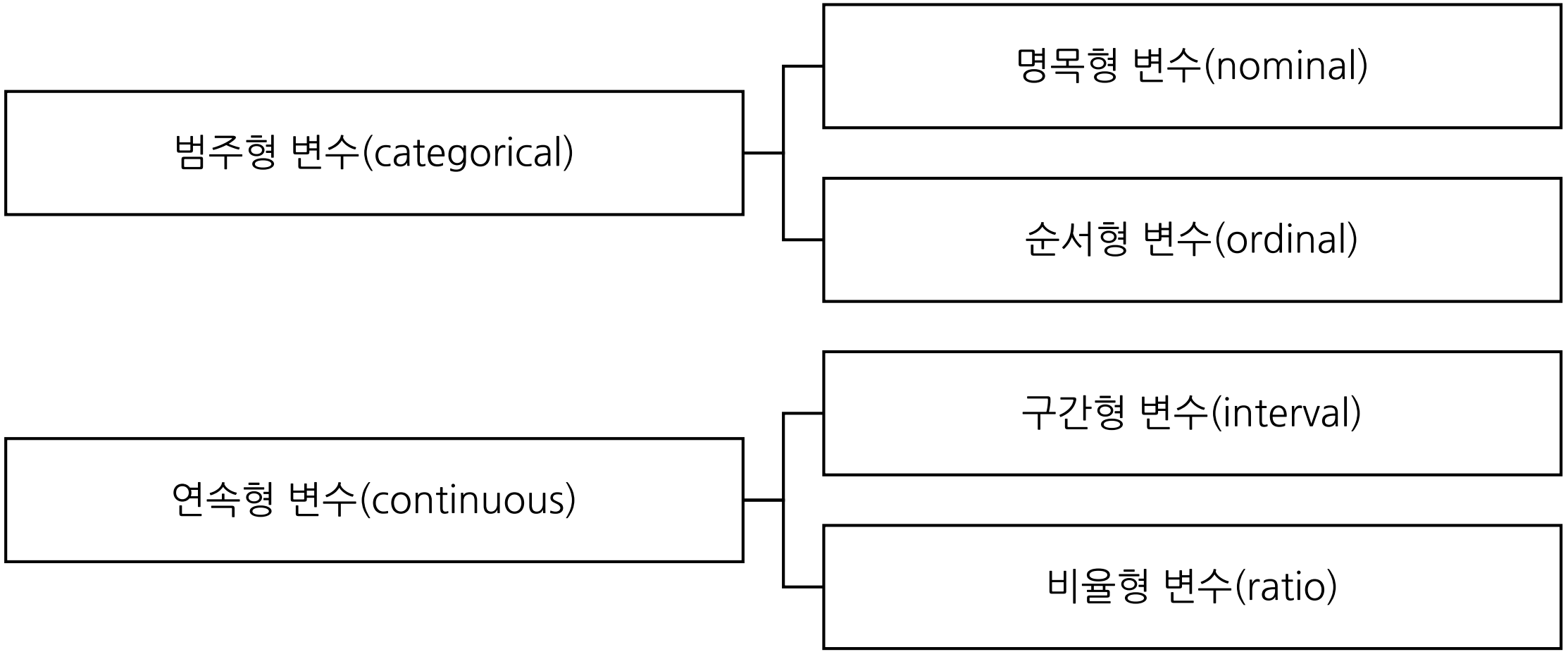
- “sex” 변수 선택하기
- df$sex
- df[,3]

(특정 개체 선택하기)

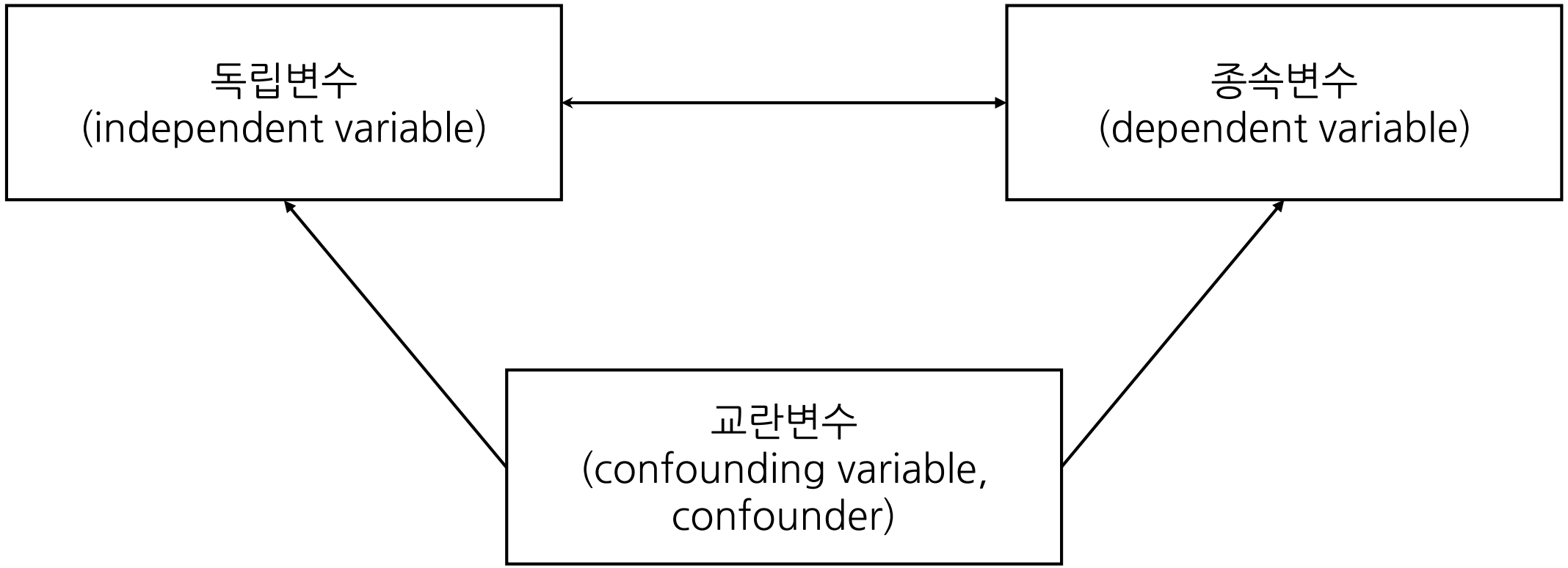
- “A901021301” 개체 선택하기
- df[2,]

2) 기본 통계이론

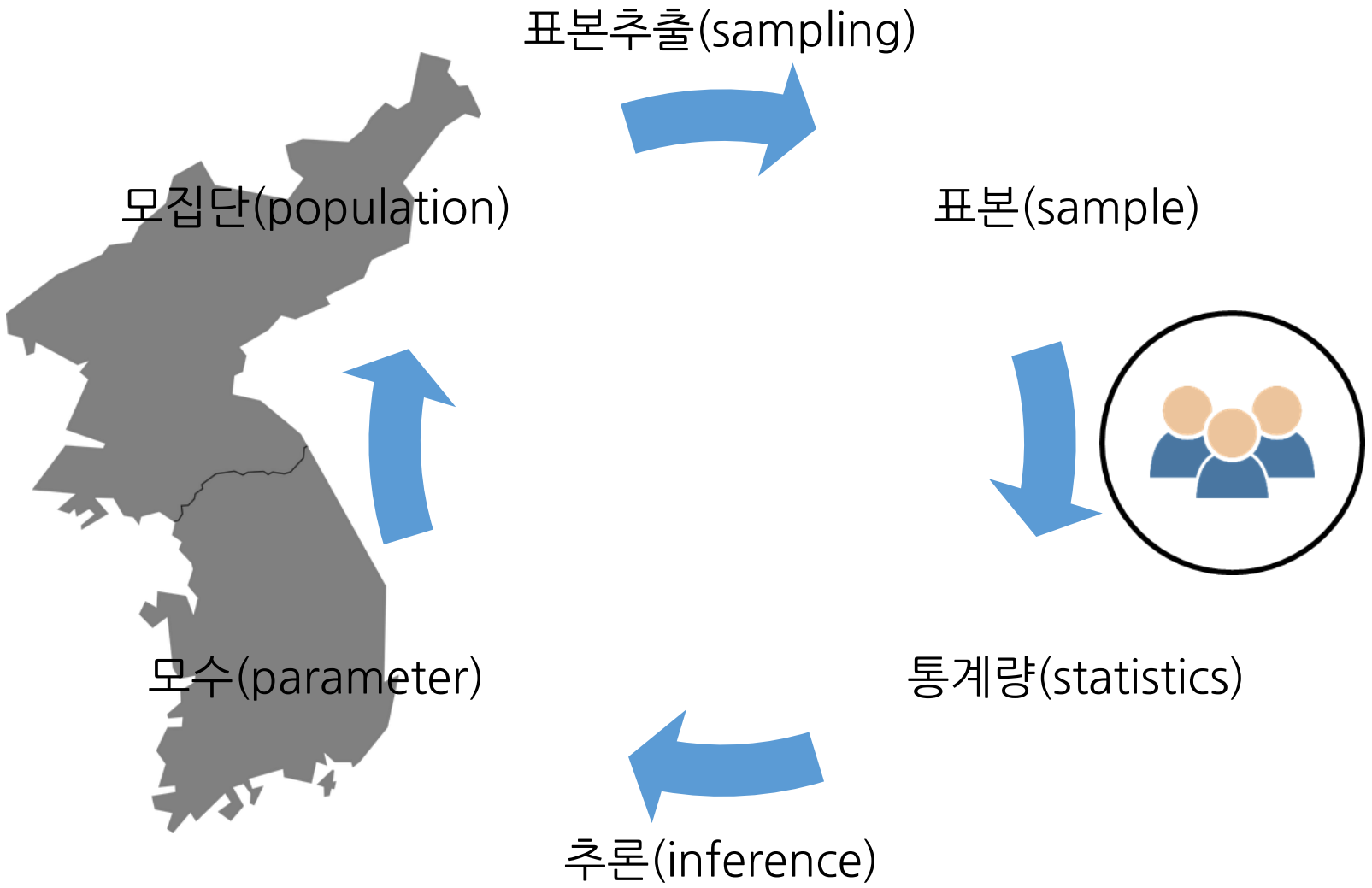
변수의 종류 1(자료의 척도에 따른 분류)



변수의 종류 2(연구가설에 따른 분류)



통계적 추론(statistical inference)



모수적 검정 vs. 비모수적 검정

모수적 검정(parametric test)

- 모집단의 분포를 모수(parameter)가 있는 확률분포로 가정하고 이에 대한 추론을 하는 표준적인 방법
- 예) Student t-test, ANOVA, Pearson 상관분석, 회귀분석

비모수적 검정(nonparametric test)

- 모집단에 대한 확률분포를 가정하지 않고 추론하는 방법
- 정규분포를 가정하지 않거나 표본 수가 매우 적은 경우
- 예) Mann-Whitney U test(Wilcoxon rank sum test), Kruskal-Wallis test, Spearman 상관분석

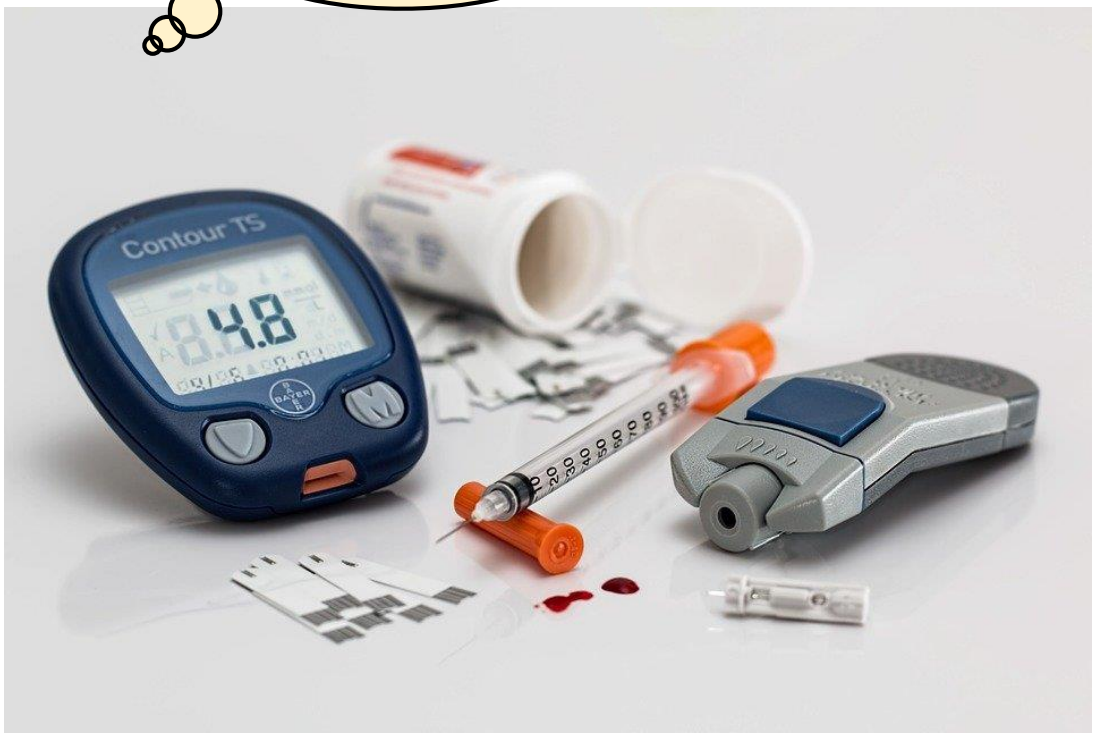
3) 통계분석을 시작하기 전에 해야 할 일

의학연구의 순서

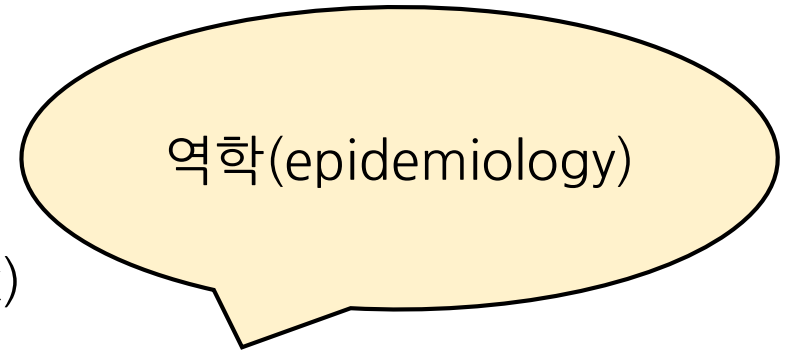
① 연구주제 탐색
② 연구가설 설정
③ 문헌고찰
④ 통계적 분석방법 설정
⑤ 데이터 수집
⑥ 데이터 정리
⑦ 통계적 분석
⑧ 결과 정리
⑨ 논문(보고서) 작성

① 연구주제 탐색

비만은 혈당을 높일까?

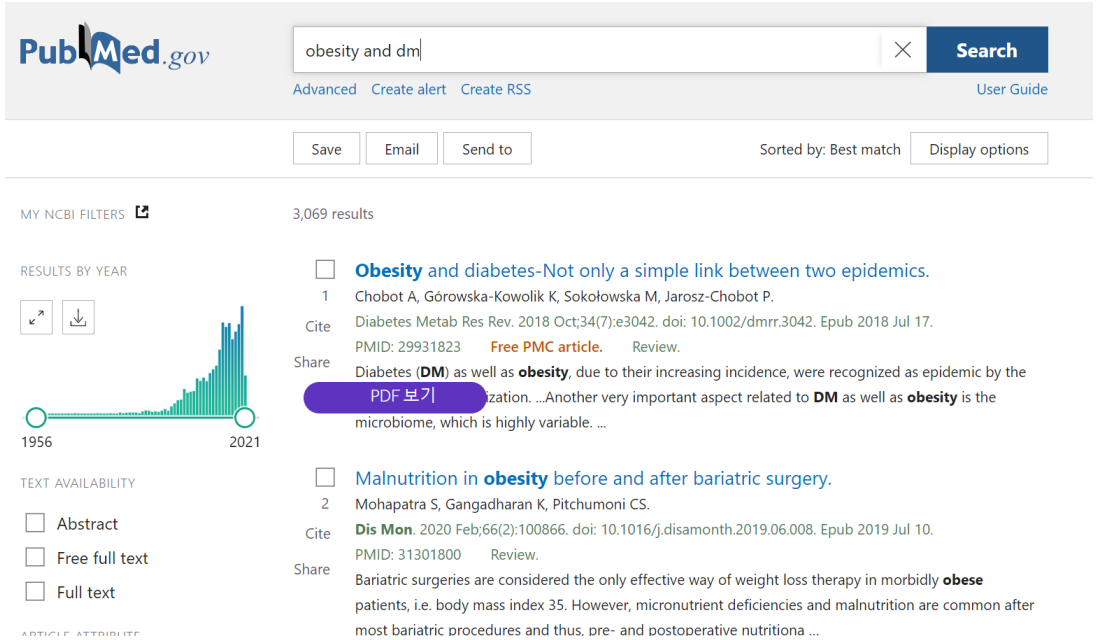


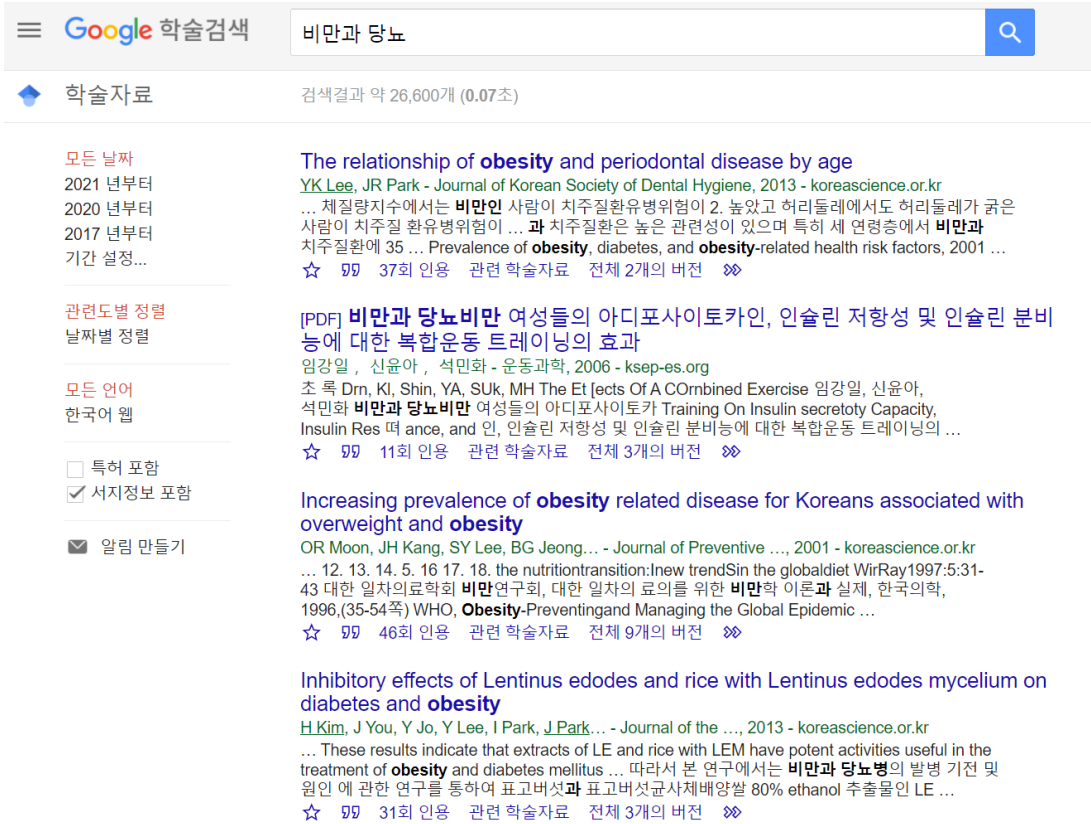
② 연구가설 설정



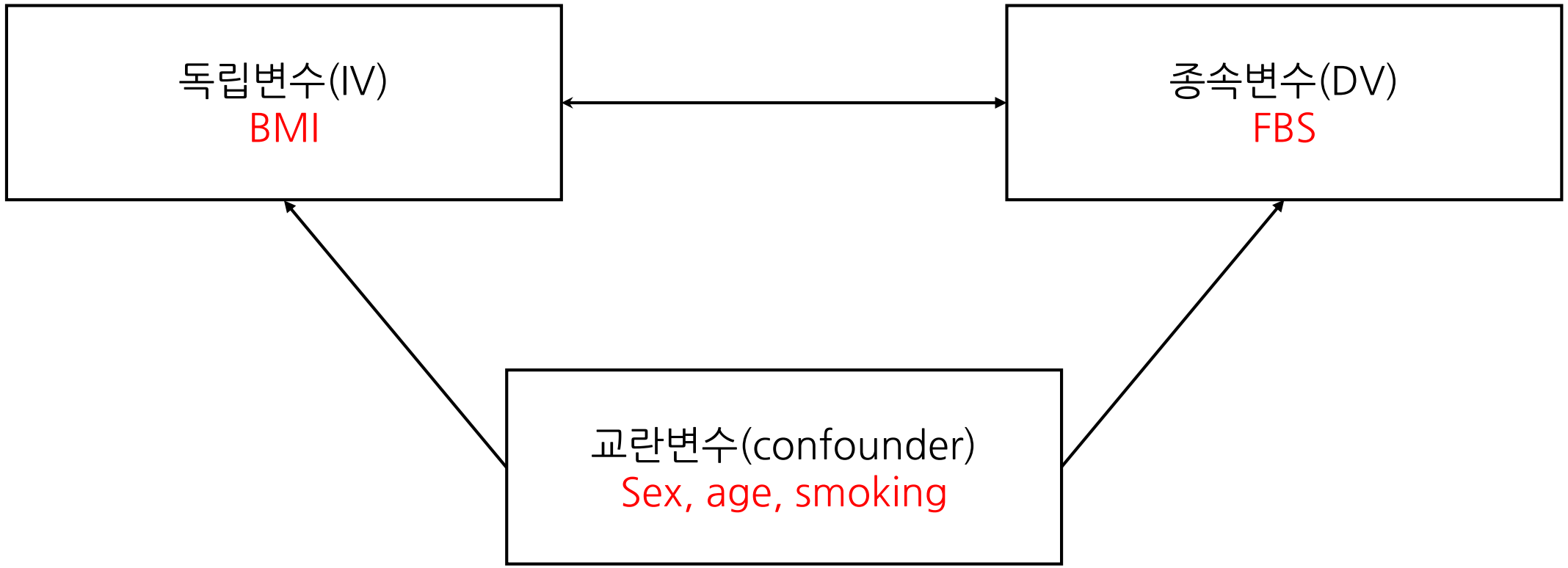
- 비만(원인) → 혈당(결과)
  - 인과성: 종단연구 ☞ 발생, RR(relative risk)
  - 연관성: 횡단연구 ☞ 유병, OR(odds ratio)
- 비만 ↔ 혈당
- 독립변수 “비만”의 정의: 체질량지수(BMI, body mass index)
- 종속변수 “혈당”의 정의: 공복혈당(FBS, fasting blood sugar)
- 연구대상 집단
- 연구가설
  - 우리나라 30세 이상 성인에서 체질량지수(BMI)의 증가와 공복혈당(FBS)의 증가는 연관성이 있다.

③ 문헌고찰

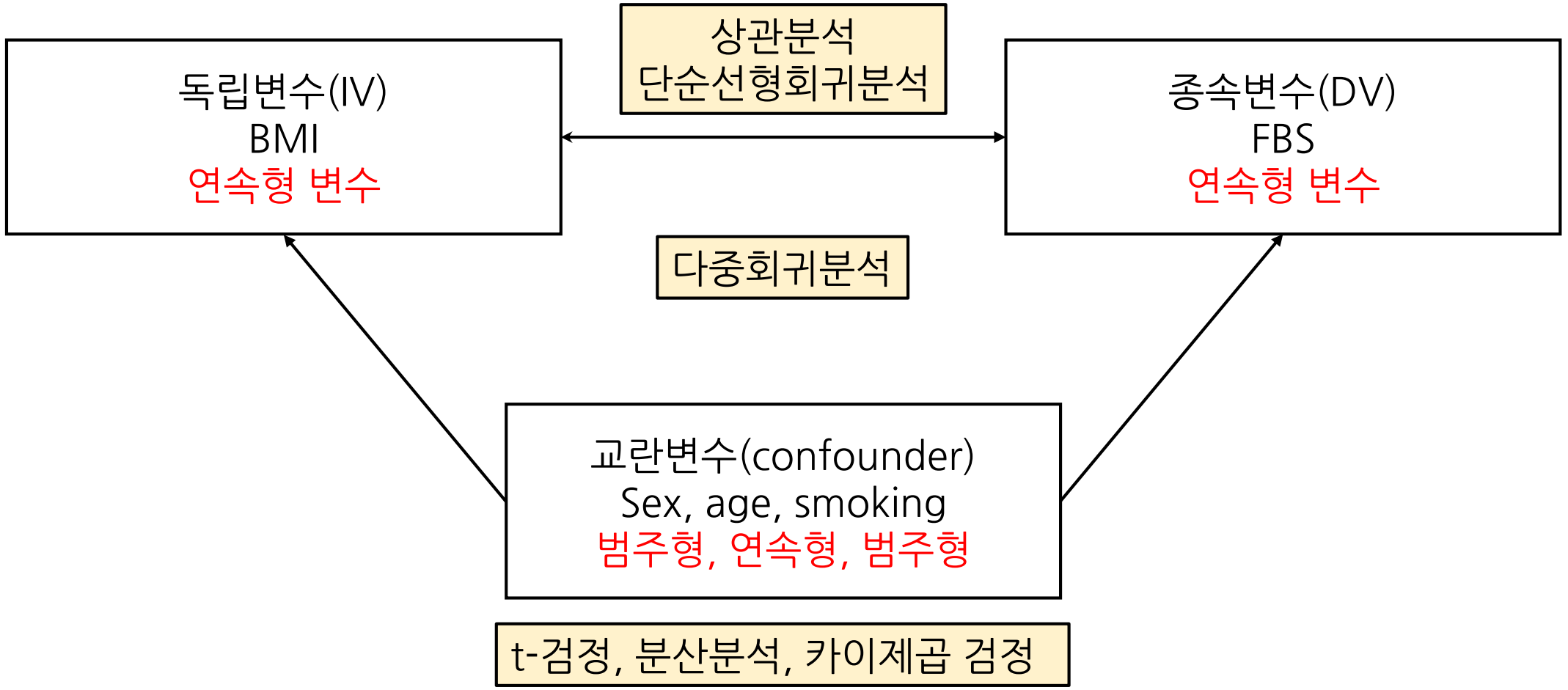

PubMed.gov
obesity and dm
Search
Advanced
Create alert
Create RSS
User Guide
Save
Email
Send to
Sorted by: Best match
Display options
MY NCBI FILTERS
3,069 results
RESULTS BY YEAR
1956
2021
TEXT AVAILABILITY
Abstract
Free full text
Full text
Obesity and diabetes-Not only a simple link between two epidemics.
Chobot A, Górowska-Kowolik K, Sokołowska M, Jarosz-Chobot P.
Cite
Diabetes Metab Res Rev. 2018 Oct;34(7):e3042. doi: 10.1002/dmrr.3042. Epub 2018 Jul 17.
PMID: 29931823
Free PMC article.
Review.
Share
Diabetes (DM) as well as obesity, due to their increasing incidence, were recognized as epidemic by the
PDF 보기
zation. ...Another very important aspect related to DM as well as obesity is the
microbiome, which is highly variable. ...
Malnutrition in obesity before and after bariatric surgery.
Mohapatra S, Gangadharan K, Pitchumoni CS.
Dis Mon. 2020 Feb;66(2):100866. doi: 10.1016/j.disamonth.2019.06.008. Epub 2019 Jul 10.
PMID: 31301800
Review.
Bariatric surgeries are considered the only effective way of weight loss therapy in morbidly obese
patients, i.e. body mass index 35. However, micronutrient deficiencies and malnutrition are common after
most bariatric procedures and thus, pre- and postoperative nutritiona ...


Google 학술검색
비만과 당뇨
학술자료
검색결과 약 26,600개 (0.07초)
모든 날짜
2021 년부터
2020 년부터
2017 년부터
기간 설정...
관련도별 정렬
날짜별 정렬
모든 언어
한국어 웹
특허 포함
서지정보 포함
알림 만들기
The relationship of obesity and periodontal disease by age
YK Lee, JR Park - Journal of Korean Society of Dental Hygiene, 2013 - koreascience.or.kr
... 체질량지수에서는 비만인 사람이 치주질환유병위험이 2. 높았고 허리둘레에서도 허리둘레가 굵은
사람이 치주질 환유병위험이 ... 과 치주질환은 높은 관련성이 있으며 특히 세 연령층에서 비만과
치주질환에 35 ... Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001 ...
37회 인용 관련 학술자료 전체 2개의 버전
[PDF] 비만과 당뇨비만 여성들의 아디포사이토카인, 인슐린 저항성 및 인슐린 분비능에 대한 복합운동 트레이닝의 효과
임강일 , 신윤아 , 석민화 - 운동과학, 2006 - ksep-es.org
초 록 Drn, Kl, Shin, YA, SUk, MH The Et [ects Of A COrnbined Exercise 임강일, 신윤아,
석민화 비만과 당뇨비만 여성들의 아디포사이토카 Training On Insulin secretoty Capacity,
Insulin Res 떠 ance, and 인, 인슐린 저항성 및 인슐린 분비능에 대한 복합운동 트레이닝의 ...
11회 인용 관련 학술자료 전체 3개의 버전
Increasing prevalence of obesity related disease for Koreans associated with overweight and obesity
OR Moon, JH Kang, SY Lee, BG Jeong... - Journal of Preventive ..., 2001 - koreascience.or.kr
... 12. 13. 14. 5. 16 17. 18. the nutritiontransition:Inew trendSin the globaldiet WirRay1997:5:31-
43 대한 일차의료학회 비만연구회, 대한 일차의 료의를 위한 비만학 이론과 실제, 한국의학,
1996,(35-54쪽) WHO, Obesity-Preventingand Managing the Global Epidemic ...
46회 인용 관련 학술자료 전체 9개의 버전
Inhibitory effects of Lentinus edodes and rice with Lentinus edodes mycelium on diabetes and obesity
H Kim, J You, Y Jo, Y Lee, I Park, J Park... - Journal of the ..., 2013 - koreascience.or.kr
... These results indicate that extracts of LE and rice with LEM have potent activities useful in the
treatment of obesity and diabetes mellitus ... 따라서 본 연구에서는 비만과 당뇨병의 발병 기전 및
원인 에 관한 연구를 통하여 표고버섯과 표고버섯균사체배양쌀 80% ethanol 추출물인 LE ...
31회 인용 관련 학술자료 전체 3개의 버전

③ 문헌고찰



④ 통계적 분석방법 설정


독립변수(IV)
BMI
연속형 변수
상관분석
단순선형회귀분석
종속변수(DV)
FBS
연속형 변수
다중회귀분석
교란변수(confounder)
Sex, age, smoking
범주형, 연속형, 범주형
t-검정, 분산분석, 카이제곱 검정

군 간 차이 검정을 위한 통계분석법

		독립적 결과변수				
		연속형		범주형		시간관찰
		정규분포	비정규분포	순서형	명목형	이분형*시간
명목형 설명변수	두 군	독립 t-검정	Wilcoxon 부호 순위합 검정 Mann-Whitney U 검정		카이제곱 검정 Fisher의 정확성검정	Log-rank 검정
	세 군 이상	일원분산분석	Kruskal-Wallis 검정		카이제곱 검정	
		의존적 결과변수				
		연속형		범주형		
		정규분포	비정규분포	순서형	명목형	
명목형 설명변수	두 군	짝지어진 t-검정	Wilcoxon 부호 순위 검정		McNemar 검정	
	두 군 이상	반복측정 분산분석	Friedman 검정		Cochran Q 검정	

자료출처: 예방의학과 공중보건학 제3판 수정증보판. pp. 238.

두 변수 간의 단순 연관성 평가를 위한 통계 분석법

			변수 1			
			연속형		범주형	
			정규분포	비정규분포	순서형	명목형
변수 2	연속형	정규분포	상관분석	Spearman 상관		분산분석
		비정규분포	상관분석	Spearman 상관		Kruskal-Wallis 검정
	명목형	순서형	Spearman 상관	Spearman 상관		Kruskal-Wallis 검정
		명목형	분산분석 Generalized linear model	Kruskal-Wallis 검정		카이제곱 검정 Fisher의 정확성검정

자료출처: 예방의학과 공중보건학 제3판 수정증보판. pp. 238.

설명변수와 결과변수 간 인과적 연관성 평가를 위한 통계분석법(교란변수 효과 보정 가능)

			독립적 결과변수					
			연속형		범주형			범주형*시간
			정규분포	비정규분포	순서형	명목형		
						3범주 이상	이분형	
설명변수	연속형	정규분포	회귀분석	치환하여 정규분포화 혹은 범주형으로 변환	순차적 로지스틱 회귀분석	명목적 로지스틱 회귀분석	로지스틱 회귀분석	Cox 회귀분석 포아송회귀분석
		비정규분포						
	명목형	순서형	회귀분석 Generalized linear model					
		명목형						

			의존적 결과변수(반복측정)				
			연속형		범주형		범주형*시간
			정규분포	비정규분포	순서형	명목형	
설명변수	연속형	정규분포	혼합모형	치환하여 정규분포화 혹은 범주형으로 변환	순차적 조건부 로지스틱 회귀분석 GEE	조건부 로지스틱 회귀분석 GEE	시간의존형 Cox 회귀분석
		비정규분포					
	명목형	순서형					
		명목형					

자료출처: 예방의학과 공중보건학 제3판 수정증보판. pp. 238.

Table 1. Demographic and clinical characteristics of study participants

Characteristics	Total (n=)
Sex	
Male	
Female	
Age (year)	
Smoking	
No	
Yes	
BMI (kg/m^2)	
FBS (mg/dl)	

BMI, body mass index; FBS, fasting blood sugar.
Data are presented as mean (SD) or n (%).

(2군 비교) Demographic and clinical characteristics of study participants by sex

Characteristics	Total (n=)	Male (n=)	Female (n=)	*p* value
Age (year)				
Smoking				
No				
Yes				
BMI (kg/m^2)				
FBS (mg/dl)				

BMI, body mass index; FBS, fasting blood sugar.
Data are presented as mean (SD) or n (%).

(3군 비교) Demographic and clinical characteristics of study participants by BMI

Characteristics	Total (n=)	Underweight (n=)	Normal (n=)	Obese (n=)	*p* value
Sex					
Male					
Female					
Age (year)					
Smoking					
No					
Yes					
FBS (mg/dl)					

BMI, body mass index; FBS, fasting blood sugar.
Data are presented as mean (SD) or n (%).

(상자그림) FBS distribution by BMI

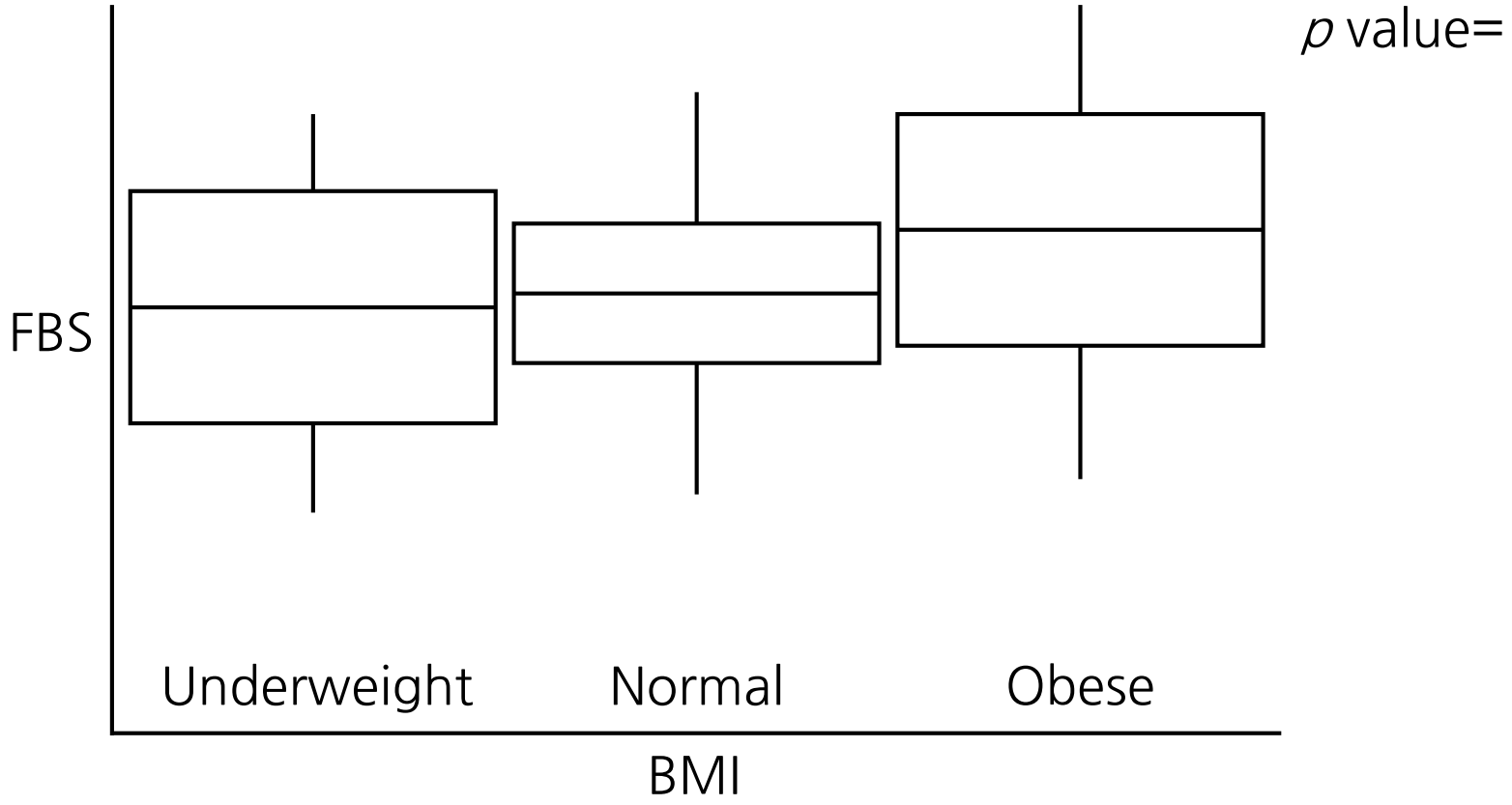


Figure 1. Correlation between BMI and FBS

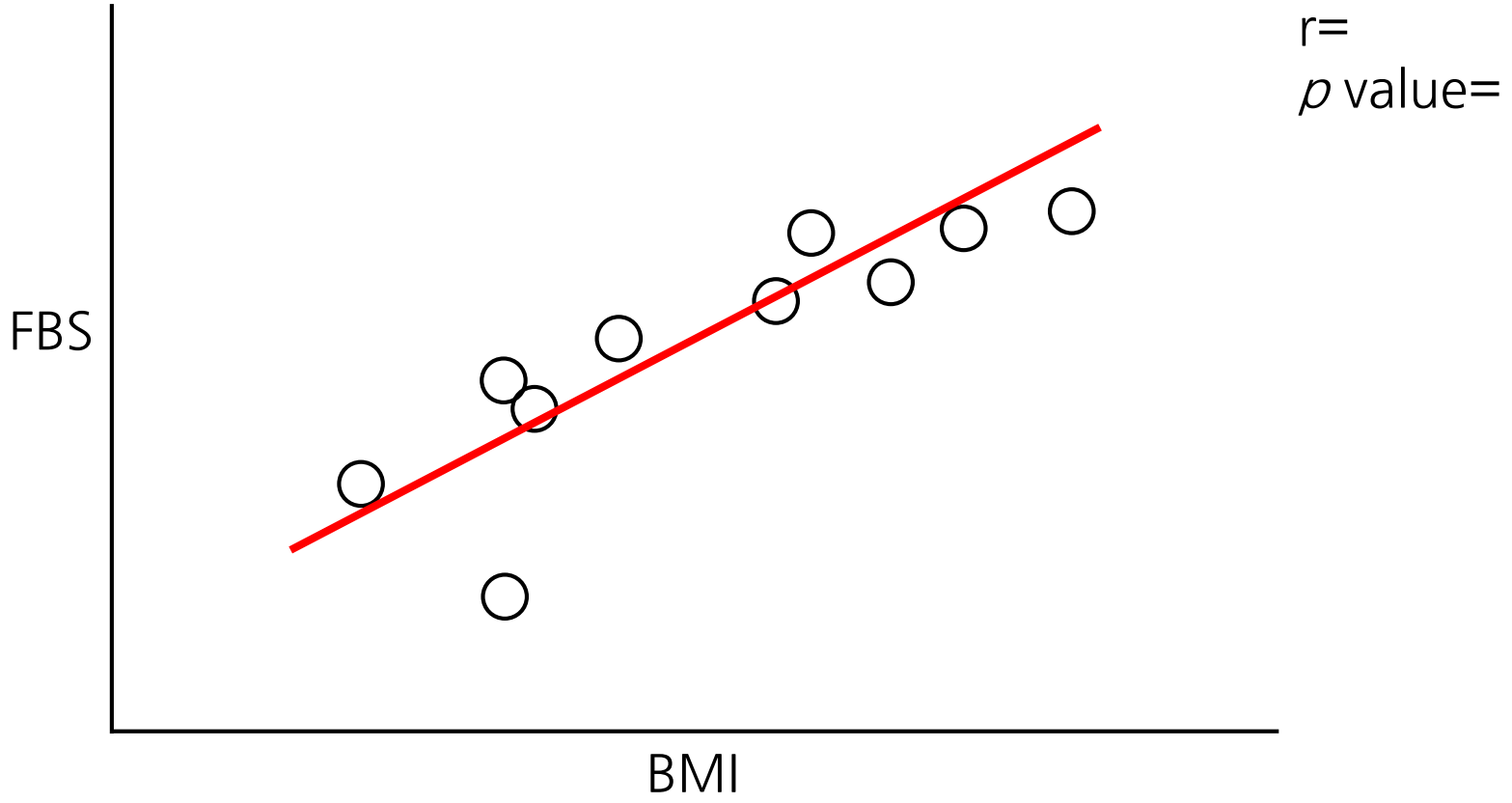


Table 2. Factors associated with FBS elevation

Factor	Univariable analysis			Multivariable analysis		
	β	Standardized β	*p* value	β	Standardized β	*p* value
Sex						
Male						
Female						
Age (year)						
Smoking						
No						
Yes						
BMI (kg/m^2)						

BMI, body mass index; FBS, fasting blood sugar.

4) 데이터 수집 및 정리

코드북 작성

변수명	변수설명	내용
id	대상자 ID	
year	조사년도	yyyy
sex	성별	1. 남자
		2. 여자
age	만나이	00세
		1~79. 1~79세
		80. 80세 이상
smk1	(성인) 평생흡연 여부	1. 5갑 미만
		2. 5갑 이상
		3. 피운 적 없음
		8. 비해당(청소년, 소아)
		9. 모름, 무응답

코드북 작성

변수명	변수설명	내용
smk2	(성인) 현재흡연 여부	1. 매일피움
		2. 가끔피움
		3. 과거엔 피웠으나, 현재 피우지 않음
		8. 비해당(smk1의 3, 8)
		9. 모름, 무응답
ht	신장	000.0cm
wt	체중	000.0kg
glu	공복혈당	000.0mg/dL

코드북 작성(new variables)

변수명	변수설명	내용
smk	현재흡연 여부	1. 과거흡연, 비흡연
		2. 현재흡연(평생 담배 5갑(100개비) 이상 피웠고 현재 담배를 피우는 경우)
bmi	체질량지수	체중(kg)/신장(m)^2
bmi_c	체질량지수 범주화	1. 저체중(0<bmi<18.5)
		2. 정상(18.5<=bmi<25)
		3. 비만(bmi>=25)
smk	현재흡연 여부	1. 과거흡연, 비흡연

데이터 수집(Excel, csv)

id	year	sex	age	smk1	smk2	ht	wt	glu
A901020501	2018	1	80	2	3	162.4	56	103
A901021301	2018	1	78	3	8	167.7	76.4	127
A901021302	2018	2	75	3	8	157.7	53.2	134
A901022601	2018	2	43	3	8	151	53.5	99
A901023303	2018	2	80	3	8	146.1	57.6	111
A901023304	2018	2	71	3	8	157.2	53.2	51
A901023501	2018	1	44	3	8	180	85.6	163
A901023502	2018	2	42	3	8	168.8	54.2	83
A901023503	2018	2	13	8	8	171.6	54.8	
A901023504	2018	1	11	8	8	155.2	38	89
A901023901	2018	1	36	3	8	172.6	69	86
A901023902	2018	2	30	3	8	153.9	54.7	104
A901023903	2018	1	2	8	8	87.5	11.3	
A901024801	2018	1	57	2	1	177.1	94.1	99

예제 데이터

- “df.csv”
- 국민건강영양조사 제7기(2016-2018) 자료 중 3차년도(2018년) 원시자료
- 성인 건강설문조사
- 7992명 대상자
- 9개 변수: 대상자 ID(“ID”), 조사년도(“year”), 성별(“sex”), 만나이(“age”), 평생흡연 여부(“BS1_1”), 현재흡연 여부(“BS3_1”), 신장(“HE_ht”), 체중(“HE_wt”), 공복혈당(“HE_glu”)
- 복합표본 가중치 고려하지 않음

국민건강영양조사(질병관리청)

국민건강영양조사(질병관리청)

- 국민건강증진법 제16조에 근거하여 국민의 건강 및 영양 상태를 파악하기 위해 매년 192개 지역의 25개 가구를 확률포본으로 추출하여 만 1세 이상 가구원 약 1만명을 조사
- 국민건강증진종합계획의 목표지표 설정 및 평가 근거자료 산출
- 흡연, 음주, 영양소섭취, 신체활동 등 건강위험행태 모니터링
- 주요 만성질환 유병률 및 관리현황(인지율, 치료율, 조절률 등) 모니터링
- 질병 및 장애에 따른 삶의 질, 활동제한, 의료이용 현황 분석
- 국가 간 비교 가능한 건강지표 산출

조사분야	조사내용 ※ 제8기 1차년도(2019년) 조사 기준
검진조사	비만, 고혈압, 당뇨병, 이상지질혈증, 간질환, 신장질환, 빈혈, 폐질환, 구강질환, 근력, 안질환, 이비인후질환
건강설문조사	가구조사, 흡연, 음주, 비만 및 체중조절, 신체활동, 이환, 의료이용, 예방접종 및 건강검진, 활동제한 및 삶의 질, 손상(사고 및 중독), 안전의식, 정신건강, 여성건강, 교육 및 경제활동, 구강건강
영양조사	식품 및 영양소 섭취현황, 식생활행태, 식이보충제, 영양지식, 식품안정성, 수유현황, 이유보충식