청소년건강패널
표본설계 및 분석방법

국립부경대 박인호

순서

1. 표본설계 – 1) 모집단

- 목표모집단: 2019년 초등학교 6학년 재학생
- 표본추출틀: KEDI 교육통계센터 교육통계 DB (2018년 자료)
- 제외대상
  - 분교, 폐ㆍ휴교(253개교), 6학년 미등록교(66개교)
  - 1학급 학교, 도서벽지, 제주지역 (1,890개교, 26,652명)
  - 규모미달 학교 (357개교, 8,640명)
- 조사모집단: 전국 3,741개교 440,194명 (2018년 4월 기준, 포함률 92.6%)

1. 표본설계 – 2) 층화 및 표본할당

- **모집단 층화: 시도 및 지역규모를 고려한 총 25개 층**
  - 지역규모는 도심화 기준으로 대도시, 중소도시, 읍면으로 분류
- **표본할당: 절충할당을 통한 전국 및 권역, 성별 추정의 효율성 고려**
  - 시도별 학교수 제곱근 비례할당
  - 지역규모별 학교수 비례할당
  - 표본층별 최소 2개 학교 이상 할당

1. 표본설계 – 3) 표본추출

- 층화다단확률추출을 적용한 학교/학급 순차적 표집
- 단계 1: 표본학교추출
  - 시도 및 지역규모 층별로 시군구, 학교규모 순 정렬후
  - 학교규모(학급수)에 비례한 계통확률추출로 표집
- 단계 2: 표본학급추출
  - 표본학교별 2개 학급을 임의추출
- 단계 3: 표본학생
  - 표본학급 내 전체 학생

2. 자료분석 –개요

- **표본가중치**
  - **표본가중치란 개별 개체가 대표하는 전체 모집단의 개체 수를 나타내는 확대배수로 아래의 구성요소를 포함**
  - **설계가중치: 수준별 표본단위 추출확률 곱의 역수**
  - **무응답조정: 횡단조사 및 추적조사 응답률(성향점수) 역수 조정**
  - **사 후 층 화: 사후층(표본층x성별)별 가중치 합 조정**

    (2019년 4월 교육통계 6학년 총 학생수 기준)
  - **가중치절사: 중앙값 기준 사분위 범위(IQR)의 4배수 기준 특이치 검출**

2. 자료분석 – 2) 표본추정

- 표본추정은 복합표본조사 내역(층,집락,가중치,조사값)을 적절히 반영해 적용
- 가중추정량

$\widehat{M}_{\cdots} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{k=1}^{m_{hi}} pw_{hik}$ 총수추정

$\bar{y} = \frac{1}{\widehat{M}_{\cdots}} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{k=1}^{m_{hi}} pw_{hik}\, y_{hik}$ 평균추정 [층 h, 학교 i, 학생 k]

pw_{hik} = 패널가중치

y_{hi} = 조사특성

n_h = 층 h 내 표본학교수

m_{hi} = 표본학교 i 내 응답학생수

2. 자료분석 – 2) 표본추정

- **선형화 분산추정식(Taylor's linearization method)**

$$v(\bar{y}) = \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi..} - \bar{e}_{h...})^2$$

$$e_{hi..} = \sum_{k=1}^{m_{hi}} pw_{hik}\ (y_{hik} - \bar{y})/\widehat{M}_{....}$$

$$\bar{e}_{h...} = \sum_{i=1}^{n_h}\ e_{hi..}/n_h$$

$1 - f_h = 1 - n_h/N_h$= 유한모집단 수정계수

N_h = 모집층 내 전체학교수

2. 자료분석 – 4) SAS 예시 (패널자료)

5년 연속 참여 대상자 선별 코딩

```
/*pwt_w1: 종단면 가중치*/
if pwt_w2="" then st2=0;else if pwt_w2^="" then st2=1;/*2차 연도 참여여부*/
if pwt_w3="" then st3=0;else if pwt_w3^="" then st3=1;/*3차 연도 참여여부*/
if pwt_w4="" then st4=0;else if pwt_w4^="" then st4=1;/*4차 연도 참여여부*/
if pwt_w5="" then st5=0;else if pwt_w5^="" then st5=1;/*5차 연도 참여여부*/
st_sum=st2+st3+st4+st5; *첫 해 제외 5년 중 참여 횟수 계산;


/*5차년도 연속 참여 여부*/
      if st_sum in (0,1,2,3,4) and st2=1 and st3=1 and st4=1 and st5=1 then st_participation5 = 2; /*연속 참여*/
else if st5 = 0 then st_participation5 = 0; /*5차년도 참여하지 않음*/
else if st5 = 1 and (st2=1 or st3=1 or st4=1 or st5=1) then st_participation5 = 1; /*불연속 참여*/
```

2. 자료분석 – 4) SAS 예시 (분석준비)

1차년도(2019) 담배제품 경험 코딩

```
/*1차년도(2019년)*/
/*tc_exp_w1: 평생 담배 제품 사용 경험률 (1): 예, (2) 아니오 ------ 해당 문항은 5차년도부터 삭제됨
- 2019~2022: 평생 담배 제품 사용 경험여부가 상위문항으로 존재
- 2022~: 각 담배제품별(일반담배(궐련), 액상형 전자담배, 궐련형 전자담배) 각각 질문하여
          하나라도 흡연 or 사용 경험이 있는 경우 평생 담배 제품 사용 경험자로 재코딩*/

/*TC_CC_LT_PF_w1: 일반담배(궐련)_평생 흡연 경험(모금 기준) (1): 예, (2) 아니오*/
if tc_exp_w1 in (1,2) then do;
  if TC_CC_LT_PF_w1 in (1) then cc_lt_w1= 1;  /*평생 일반 담배 제품 흡연 경험이 있는 경우*/
    else if TC_EXP_w1  in (2) or TC_CC_LT_PF_w1 in (2) then cc_lt_w1 =0; /*평생 담배 제품 사용 경험X  or 일반담배 (모금) 경험X*/

/*TC_EC_LT_w1: 액상형_평생 흡연 여부(모금 기준)
- (1): 예, (2)아니오, (3)액상형 전자담배(베이핑)를 모름*/
if TC_EC_LT_w1 in (1) then ec_lt_w1= 1;
    else if TC_EXP_w1  in (2) or TC_EC_LT_w1 in (2,3) then ec_lt_w1 =0;/*평생 액상형 전자담배 사용 경험이 있는 경우*/

/*TC_HNB_LT_PF_w1: 궐련형 전자담배_평생 흡연 여부(모금 기준)
- (1): 예, (2)아니오, (3)궐련형 전자담배를 모름*/
if TC_HNB_LT_PF_w1 in (1) then htp_lt_w1 = 1;
    else if TC_EXP_w1 in (2) or TC_HNB_LT_PF_w1 in (2,3)   then htp_lt_w1 =0;/*평생 궐련형 전자담배 사용 경험이 있는 경우*/


/*1차년도(2019) 담배제품 사용 경험 여부*/
if cc_lt_w1=1 or ec_lt_w1=1 or htp_lt_w1=1 then tc_lt_w1=1; *1차년도에서 흡연 경험이 있음;
    else if cc_lt_w1=0 and ec_lt_w1=0 and htp_lt_w1=0 then tc_lt_w1=0; end; /*1차년도에서 담배제품 무경험자*/
```

2. 자료분석 – 4) SAS 예시 (자료준비)

담배제품 현재 사용자 코딩

```
/*-----------------------------------------------
            담배제품 현재사용여부

# TC_CC_DAYS_w4: 일반담배(궐련)_월간 흡연 빈도
1: 최근 30일 동안 없다
2: 월 1-2일
3:  월 3-5일
4: 월 6-9일
5: 월 10-19일
6: 월 20-29일
7: 매일
-----------------------------------------------*/

    if tc_exp_w4 in (1,2) then do;

    if  TC_CC_DAYS_w4 in (2:7) then cc_current_w4=1;else cc_current_w4=0;
    if  TC_EC_DAYS_w4 in (2:7) then ec_current_w4=1;else ec_current_w4=0;
    if  TC_hnb_DAYS_w4 in (2:7) then hnb_current_w4=1;else hnb_current_w4=0;
 if cc_current_w4=1 or ec_current_w4=1 or hnb_current_w4=1 then all_current_w4=1;
 else if cc_current_w4=0 and ec_current_w4=0 and hnb_current_w4=0 then all_current_w4=0;end;
```

2. 자료분석 – 4) SAS 예시 (자료준비)

5차년도(2023) 담배제품 경험 코딩

```
/*평생 일반담배(궐련) 흡연 경험(모금) 여부*/
  if TC_CC_LT_PF_w5 in (1) then cc_lt_w5= 1; *유경험;
    else if TC_EXP_w5  in (2) or TC_CC_LT_PF_w5 in (2) then cc_lt_w5 =0; *무경험;

/*평생 액상형 전자담배 사용 경험 여부*/
  if TC_EC_LT_w5 in (1) then ec_lt_w5= 1; *유경험;
    else if TC_EXP_w5  in (2) or TC_EC_LT_w5 in (2,3) then ec_lt_w5 =0;*무경험;

/*평생 궐련형 전자담배 사용 경험 여부*/
if TC_HNB_LT_PF_w5 in (1) then htp_lt_w5 = 1; *유경험;
    else if TC_EXP_w5 in (2) or TC_HNB_LT_PF_w5 in (2,3)   then htp_lt_w5 =0;*무경험;

/*평생 담배제품 경험 여부*/
if cc_lt_w5=1 or ec_lt_w5=1 or htp_lt_w5=1 then tc_lt_w5=1; *유경험;
     else if cc_lt_w5=0 and ec_lt_w5=0 and htp_lt_w5=0 then tc_lt_w5=0;*무경험;
```

2. 자료분석 – 4) SAS 예시 (담배제품 현재 사용 prevalence)

현재 사용률

```
/*2020년*/
proc surveyfreq data=tc_inci total=st_sch  nomcar;
strata st;
cluster SchID; *학교번호*;
weight pwt_w2; *2차년도 종단면 가중치;
tables all_current_w2/deff;
run;
```

SAS 시스템

The SURVEYFREQ Procedure

Data Summary	
Number of Strata	25
Number of Clusters	260
Number of Observations	5051
Number of Observations Used	4702
Number of Obs with Nonpositive Weights	349
Sum of Weights	468218.54

Variance Estimation	
Method	Taylor Series
Missing Values	NOMCAR

Table of all_current_w2

all_current_w2	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Design Effect
0	4699	468070	6750	99.9683	0.0181	0.5201
1	3	148.42754	84.60136	0.0317	0.0181	0.5201
Total	4702	468219	6752	100.000		

***all_current_w2**
: 2차년도(2020) 현재 흡연 여부
0 (현재 비흡연)
1 (현재 흡연 중)

2. 자료분석 - 4. SAS 예시 (자료준비)

담배제품 시작 연도

```
/*담배 무관 시작 응답 연도 코딩*/
if st_participation5=2 then do; *5차연도까지 연속으로 참여한 대상자;
  if tc_lt_w1=1 or tc_lt_w2=1 or tc_lt_w3=1 or tc_lt_w4=1 or tc_lt_w5=1then tc_lt=1; *한번이라도 흡연경험O;
  else if tc_lt_w1=0 and tc_lt_w2=0 and tc_lt_w3=0 and tc_lt_w4=0  and tc_lt_w5=0  then tc_lt=0; *평생 담배제품 무경험자;

   if tc_lt=1 then do; *흡연 유경험 대상자;

            if tc_lt_w1=1 then tc_start_year=1; *흡연을 시작한 연도;
       else if tc_lt_w2=1 then tc_start_year=2;
       else if tc_lt_w3=1 then tc_start_year=3;
       else if tc_lt_w4=1 then tc_start_year=4;
       else if tc_lt_w5=1 then tc_start_year=5; end;

 if tc_lt=0 then tc_start_year=0; *흡연 무경험자;

   /*1->2차연도 경험*/
         if tc_start_year in (1) then tc_start_year2=0;/*1차연도 흡연->분석 제외*/
  else if tc_start_year in (0,3,4,5) then tc_start_year2=1;/*2차연도 비흡연*/
  else if tc_start_year in (2) then tc_start_year2=2;/*2차연도 흡연*/

  /*2->3차연도 경험*/
         if tc_start_year in (1,2) then tc_start_year3=0;/*1,2차연도 흡연->분석 제외*/
  else if tc_start_year in (0,4,5) then tc_start_year3=1;/*3차연도 비흡연*/
  else if tc_start_year in (3) then tc_start_year3=2; end;  /*3차연도 흡연*/
```

2. 자료분석 - 4. SAS 예시 (담배제품 신규 사용 incidence)

연차별 신규 흡연 발생률

```
/*2019 -> 2020년*/
proc surveyfreq data=tc_inci total=st_sch  nomcar;
strata st;
cluster SchID; *학교번호*;
weight pwt_w5; *5차년도 종단면 가중치;
tables st_participation5 * (tc_start_year2 ) /row deff;
where tc_start_year2 in (1, 2);
run;
```

SAS 시스템

The SURVEYFREQ Procedure

Data Summary	
Number of Strata	25
Number of Clusters	260
Number of Observations	3976
Sum of Weights	465923.618

Variance Estimation	
Method	Taylor Series
Missing Values	NOMCAR

Table of st_participation5 by tc_start_year2

st_participation5	tc_start_year2	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Design Effect	Row Percent	Std Err of Row Percent
2	1	3960	464533	8629	99.7016	0.0730	0.7651	99.7016	0.0730
	2	16	1390	339.48392	0.2984	0.0730	0.7651	0.2984	0.0730
	Total	3976	465924	8627	100.000			100.000	

***st_participation5=2**
: 5차년도까지 연속 참여 대상자

***tc_start_year2**
: 2차년도 흡연 신규 발생
1 (2차년도 비흡연)
2 (2차년도 신규흡연)

2. 자료분석 – 4) SAS 예시 (로지스틱 분석, 성별)

빈도와 가중치

```
proc surveyfreq data=tc_inci total=st_sch nomcar;
strata st;
cluster SchID; *학교번호*;
weight pwt_w5;
tables sex*tc_lt/row chisq;
where tc_lt in (0,1) and tc_lt_w1=0 ;
run;
```

tc_lt
0: 흡연 경험 없음
1: 흡연 경험 있음

tc_lt_w1
0: 1차년도 흡연 경험 없음

SEX
1: 남성
2: 여성

Table of SEX by tc_lt

SEX	tc_lt	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
1	0	1901	216933	5608	46.5597	0.8897	90.7069	0.6993
	1	192	22225	1831	4.7701	0.3707	9.2931	0.6993
	Total	2093	239158	6129	51.3299	0.9073	100.000	
2	0	1783	215299	5956	46.2090	0.9291	94.9433	0.6173
	1	100	11467	1402	2.4611	0.3012	5.0567	0.6173
	Total	1883	226766	5952	48.6701	0.9073	100.000	
Total	0	3684	432232	8182	92.7688	0.4759		
	1	292	33692	2337	7.2312	0.4759		
	Total	3976	465924	8627	100.000			

OR(95% CI)

```
proc surveylogistic data=tc_inci total=st_sch nomcar;
strata st;
cluster SchID; *학교번호;
weight pwt_w5;
class sex(ref="2") /ref=first param=ref;
model tc_lt (ref='0')=sex ; *비흡연유지 학생 기준;
where tc_lt in (0,1) and tc_lt_w1=0 ;
run;
```

Odds Ratio Estimates

Effect	Point Estimate	95% Confidence Limits	
SEX 1 vs 2	1.924	1.426	2.594

NOTE: The degrees of freedom in computing the confidence limits is 235.

2. 자료분석 – 5) R 예시 (패널자료)

5년 연속 참여 대상자 선별 코딩

```r
panel <- st12345 |>
  mutate(                                                    # 연도별 참여여부
    st2 = if_else(is.na(pwt_w2) | pwt_w2 == "", 0L, 1L),     # 0L, 1L 정수형 반환
    st3 = if_else(is.na(pwt_w3) | pwt_w3 == "", 0L, 1L),
    st4 = if_else(is.na(pwt_w4) | pwt_w4 == "", 0L, 1L),
    st5 = if_else(is.na(pwt_w5) | pwt_w5 == "", 0L, 1L),
    st_sum = st2 + st3 + st4 + st5,

    # 5차년도 연속 참여 여부
    st_participation5 = case_when(
      st2 == 1 & st3 == 1 & st4 == 1 & st5 == 1 ~ 2L,   # 연속 참여
      st5 == 0                                  ~ 0L,   # 5차 미참여
      st5 == 1                                  ~ 1L    # 불연속 참여
    )
  )
```

2. 자료분석 – 5) R 예시 (분석준비)

1차년도(2019) 담배제품 경험 코딩

```
mutate(

  # ── 1차년도 (2019) ──────────────────────────────────────
  cc_lt_w1 = case_when(
    tc_cc_lt_pf_w1 == 1                                ~ 1L,
    tc_exp_w1 == 2 | tc_cc_lt_pf_w1 == 2               ~ 0L
  ),
  ec_lt_w1 = case_when(
    tc_ec_lt_w1 == 1                                    ~ 1L,
    tc_exp_w1 == 2 | tc_ec_lt_w1 %in% c(2, 3)          ~ 0L
  ),
  htp_lt_w1 = case_when(
    tc_hnb_lt_pf_w1 == 1                                ~ 1L,
    tc_exp_w1 == 2 | tc_hnb_lt_pf_w1 %in% c(2, 3)      ~ 0L
  ),
  tc_lt_w1 = case_when(
    cc_lt_w1 == 1 | ec_lt_w1 == 1 | htp_lt_w1 == 1     ~ 1L,
    cc_lt_w1 == 0 & ec_lt_w1 == 0 & htp_lt_w1 == 0     ~ 0L
  ),
  cc_current_w1  = if_else(tc_cc_days_w1  %in% 2:7, 1L, 0L),
  ec_current_w1  = if_else(tc_ec_days_w1  %in% 2:7, 1L, 0L),
  hnb_current_w1 = if_else(tc_hnb_days_w1 %in% 2:7, 1L, 0L),
  all_current_w1 = case_when(
    cc_current_w1 == 1 | ec_current_w1 == 1 | hnb_current_w1 == 1 ~ 1L,
    cc_current_w1 == 0 & ec_current_w1 == 0 & hnb_current_w1 == 0 ~ 0L
  ),
```

2. 자료분석 – 4) SAS 예시 (자료준비)

담배제품 현재 사용자 코딩

```
# 최초 현재 흡연 연도 (tc_current_y)
tc_current_y = case_when(
  all_current_w5 == 1 ~ 5L,
  all_current_w4 == 1 ~ 4L,
  all_current_w3 == 1 ~ 3L,
  all_current_w2 == 1 ~ 2L,
  all_current_w1 == 1 ~ 1L,
  TRUE                ~ NA_integer_
)
```

2. 자료분석 – 5) R 예시 (자료준비)

5차년도(2023) 담배제품 경험 코딩

```r
# ── 5차년도 (2023) ──────────────────────────────────────
# tc_exp_w5: 각 제품별 경험 문항으로 재구성
tc_exp_w5 = if_else(
  tc_cc_lt_pf_w5 == 1 | tc_ec_lt_w5 == 1 | tc_hnb_lt_pf_w5 == 1,
  1L, 2L
),
cc_lt_w5 = case_when(
  tc_cc_lt_pf_w5 == 1                              ~ 1L,
  tc_exp_w5 == 2 | tc_cc_lt_pf_w5 == 2             ~ 0L
),
ec_lt_w5 = case_when(
  tc_ec_lt_w5 == 1                                  ~ 1L,
  tc_exp_w5 == 2 | tc_ec_lt_w5 %in% c(2, 3)         ~ 0L
),
htp_lt_w5 = case_when(
  tc_hnb_lt_pf_w5 == 1                              ~ 1L,
  tc_exp_w5 == 2 | tc_hnb_lt_pf_w5 %in% c(2, 3)     ~ 0L
),
tc_lt_w5 = case_when(
  cc_lt_w5 == 1 | ec_lt_w5 == 1 | htp_lt_w5 == 1    ~ 1L,
  cc_lt_w5 == 0 & ec_lt_w5 == 0 & htp_lt_w5 == 0    ~ 0L
),
```

2. 자료분석 – 4) R 예시 (자료준비)

복합표본 설계 객체 생성

```r
make_design <- function(data, weight_var) {
  data <- data |>
    filter(!is.na(.data[[weight_var]]),           # 가중치 NA 제거
           as.numeric(.data[[weight_var]]) > 0)   # 0 이하 가중치 제거

  svydesign(
    ids     = ~SchID,     # 집락정보
    strata  = ~st,        # 층정보
    weights = as.formula(paste0("~", weight_var)), # 표본가중치
    fpc     = ~nopsu,     # 층별 집락총수
    data    = data,
    nest    = TRUE
  )
}
```

2. 자료분석 – 5) R 예시 (담배제품 현재 사용 prevalence)

현재 사용률

```r
des_w2 <- make_design(tc_inci, "pwt_w2")

# 가중 빈도 (weighted count)
svytable(~all_current_w2, design = des_w2)

# 비율 + 표준오차 (SE)
svymean(~factor(all_current_w2), design = des_w2)

# 비율 + 표준오차 + 95% 신뢰구간
confint(svymean(~factor(all_current_w2), design = des_w2))

# 비율 + SE + 95% CI + DEFF 한번에 출력
result_prev_w2 <- svymean(~factor(all_current_w2), design = des_w2, deff = TRUE)
data.frame(
  category   = c("비흡연(0)", "현재흡연(1)"),
  proportion = coef(result_prev_w2),
  SE         = SE(result_prev_w2),
  CI_lower   = confint(result_prev_w2)[, 1],
  CI_upper   = confint(result_prev_w2)[, 2],
  DEFF       = deff(result_prev_w2)
)
```

	category	proportion	SE	CI_lower	CI_upper	DEFF
factor(all_current_w2)0	비흡연(0)	0.9996829952	0.0001806161	9.993290e-01	1.0000369962	0.4888305
factor(all_current_w2)1	현재흡연(1)	0.0003170048	0.0001806161	-3.699621e-05	0.0006710059	0.4888305

2. 자료분석 – 5) R 예시 (자료준비)

담배제품 시작 연도

```
mutate(

  # 5차까지 연속 참여자만 처리
  tc_lt = case_when(
    st_participation5 != 2 ~ NA_integer_,
    tc_lt_w1 == 1 | tc_lt_w2 == 1 | tc_lt_w3 == 1 |
      tc_lt_w4 == 1 | tc_lt_w5 == 1                ~ 1L,
    tc_lt_w1 == 0 & tc_lt_w2 == 0 & tc_lt_w3 == 0 &
      tc_lt_w4 == 0 & tc_lt_w5 == 0                ~ 0L
  ),

  # 흡연 시작 연도
  tc_start_year = case_when(
    st_participation5 != 2 ~ NA_integer_,
    tc_lt == 0             ~ 0L,
    tc_lt_w1 == 1          ~ 1L,
    tc_lt_w2 == 1          ~ 2L,
    tc_lt_w3 == 1          ~ 3L,
    tc_lt_w4 == 1          ~ 4L,
    tc_lt_w5 == 1          ~ 5L
  ),

  # 1→2차 신규 경험
  # 0: 1차 흡연(제외), 1: 2차 비흡연, 2: 2차 흡연
  tc_start_year2 = case_when(
    tc_start_year == 1           ~ 0L,
    tc_start_year %in% c(0,3,4,5) ~ 1L,
    tc_start_year == 2           ~ 2L
  ),

  # 2→3차 신규 경험
  tc_start_year3 = case_when(
    tc_start_year %in% c(1, 2)    ~ 0L,
    tc_start_year %in% c(0, 4, 5) ~ 1L,
    tc_start_year == 3            ~ 2L
  ),
```

2. 자료분석 – 5) R 예시 (담배제품 신규 사용 incidence)

연차별 신규 흡연 발생률

```r
des_w5 <- make_design(tc_inci, "pwt_w5")

# 분석 대상: tc_start_year2 in (1, 2)
sub_inci <- subset(des_w5, tc_start_year2 %in% c(1, 2))

# 가중 빈도 (weighted count)
svytable(~st_participation5 + tc_start_year2, design = sub_inci)

# 행별(st_participation5) 신규 흡연(tc_start_year2==2) 비율 + SE + 95% CI + DEFF
result_inci <- svyby(
  formula = ~I(tc_start_year2 == 2),    # 신규 흡연 여부
  by      = ~st_participation5,         # 행 변수
  design  = sub_inci,
  FUN     = svymean,
  deff    = TRUE,
  vartype = c("se", "ci")               # SE + 95% CI 동시 출력
)

# TRUE(신규 흡연) 컬럼만 추출하여 정리
#    컬럼 순서: [1]=by변수, [2]=FALSE비율, [3]=TRUE비율,
#               [4]=FALSE_SE, [5]=TRUE_SE, [6]=FALSE_CIl, [7]=TRUE_CIl,
#               [8]=FALSE_CIu, [9]=TRUE_CIu, [10]=FALSE_DEFF, [11]=TRUE_DEFF

data.frame(
  st_participation5 = result_inci[[1]],
  proportion        = result_inci[[3]],    # TRUE 비율
  SE                = result_inci[[5]],    # TRUE SE
  CI_lower          = result_inci[[7]],    # TRUE CI lower
  CI_upper          = result_inci[[9]],    # TRUE CI upper
  DEFF              = result_inci[[11]]    # TRUE DEFF
)
```

```
  st_participation5  proportion           SE    CI_lower   CI_upper      DEFF
1                 2 0.002983867 0.0007299232 0.001553244 0.00441449 0.7180121
```

2. 자료분석 – 5) R 예시 (로지스틱 분석, 성별)

빈도와 가중치

```r
# 분석 대상: tc_lt %in% (0,1) & tc_lt_w1 == 0
sub_lgst_w2 <- subset(des_w2, tc_lt %in% c(0, 1) & tc_lt_w1 == 0)
sub_lgst_w5 <- subset(des_w5, tc_lt %in% c(0, 1) & tc_lt_w1 == 0)

# --- 빈도와 가중치 (%) ---

# 가중 빈도 교차표
svytable(~sex + tc_lt, design = sub_lgst_w2)

# 성별(sex)별 흡연경험(tc_lt==1) 비율 + SE + 95% CI + DEFF
result_freq <- svyby(
  formula = ~I(tc_lt == 1),
  by      = ~sex,
  design  = sub_lgst_w2,
  FUN     = svymean,
  deff    = TRUE,
  vartype = c("se", "ci")
)

# TRUE 컬럼명으로 직접 선택하여 출력
true_cols <- grep("TRUE$", names(result_freq), value = TRUE)  # TRUE 컬럼만
result_freq[, c("sex", true_cols)]

# TRUE(흡연 경험) 컬럼만 추출하여 정리
#    컬럼 순서: [1]=by변수, [2]=FALSE비율, [3]=TRUE비율,
#               [4]=FALSE_SE, [5]=TRUE_SE, [6]=FALSE_CIl, [7]=TRUE_CIl,
#               [8]=FALSE_CIu, [9]=TRUE_CIu, [10]=FALSE_DEFF, [11]=TRUE_DEFF
data.frame(
  sex        = result_freq[[1]],
  proportion = result_freq[[3]],    # TRUE 비율
  SE         = result_freq[[5]],    # TRUE SE
  CI_lower   = result_freq[[7]],    # TRUE CI lower
  CI_upper   = result_freq[[9]],    # TRUE CI upper
  DEFF       = result_freq[[11]]    # TRUE DEFF
)
```

```
  sex proportion       SE    CI_lower    CI_upper       DEFF
1   1   0.091631 0.908369 0.006756846 0.006756846 0.07838783
2   2   0.050637 0.949363 0.005894500 0.005894500 0.03908399
```

2. 자료분석 – 5) R 예시 (로지스틱 분석, 성별)

OR(95% CI)

```r
tc_inci_lgst <- tc_inci |>
  filter(tc_lt %in% c(0, 1), tc_lt_w1 == 0,
         !is.na(pwt_w5), as.numeric(pwt_w5) > 0) |>   # 가중치 NA·0 제거
  mutate(
    tc_lt_f = factor(tc_lt, levels = c(0, 1)),        # ref = 0 (비흡연 유지)
    sex_f   = relevel(factor(sex), ref = "2")         # ref = 2 (여학생)
  )

des_lgst <- svydesign(
  ids     = ~SchID,
  strata  = ~st,
  weights = ~pwt_w5,
  fpc     = ~nopsu,
  data    = tc_inci_lgst,
  nest    = TRUE
)

fit_lgst <- svyglm(
  tc_lt_f ~ sex_f,
  design = des_lgst,
  family = quasibinomial(link = "logit")
)

summary(fit_lgst)

# OR 및 95% CI 출력
exp(cbind(OR = coef(fit_lgst), confint(fit_lgst)))
```

```
Call:
svyglm(formula = tc_lt_f ~ sex_f, design = des_lgst, family = quasibinomial(link = "logit"))

Survey design:
svydesign(ids = ~SchID, strata = ~st, weights = ~pwt_w5, fpc = ~nopsu,
    data = tc_inci_lgst, nest = TRUE)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.9326     0.1286 -22.808  < 2e-16 ***
sex_f1        0.6542     0.1518   4.311  2.4e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.000252)

Number of Fisher Scoring iterations: 5
```

```
                       OR      2.5 %     97.5 %
(Intercept) 0.05326054 0.04134202 0.06861507
sex_f1      1.92359212 1.42645229 2.59399256
```

2. 자료분석 – 6) 주의사항

- **결과 해석 시 유의사항**
  - **표본가중치 미사용 ➔ 편향 가능 (가중치는 대표성 회복 장치)**
  - **집락 미사용 ➔ (주로) 표본오차 과소추정 (집락 사용에 따른 정도수준 상실 미반영)**
  - **희귀특성(rare trait)은 상대표준오차 대신 신뢰구간 중심 해석 필요 (Parker et al., 2017*)**

*** Parker et al. (2017). National Center for Health Statistics**
data presentation standards for proportions, NCHS, Vital Health Stat. 2 (175).

감사합니다.

A. 참고: 표본가중치 상세논의

- **설계가중치 및 무응답 조정**
  - **표본학교 설계가중치**

$$w_{hi} = \frac{1}{\text{학교추출률}} = \frac{\text{표본층 내 총학급수}}{(\text{표본학교수} \times \text{학교 내 총학급수})}$$

  - **표본학급 및 학생 추출확률**

$$p_{hi} = \frac{\text{학교 내 표본학급수}}{\text{학교 내 총학급수}} \;\&\; p_{hijk} = 1 \text{ (전수)}$$

  - **표본학생 무응답 조정 가중치**

$$w_{hijk} = w_{hi} p_{hij}^{-1} p_{hijk}^{-1} = \frac{\text{표본층 내 총학급수} \times \text{학급 내 총학생수}}{\text{표본학교수} \times \text{표본학급수} \times \text{학급 내 응답학생수}}$$

A. 참고: 표본가중치 상세논의

- **사후층화 및 가중치절사**
  - **사후층(표본층*성별)별 2019년 4월 기준 교육통계 DB 자료 보정**

$$w_{hik}^{PS} = w_{hik} \frac{\text{사후층 } 2019 \text{ 교육통계 총합 } (M_\gamma)}{\text{사후층 조정전 가중합 } (\widehat{M}_\gamma)}$$

$$\Rightarrow \sum_{(hik)\in s_\gamma} w_{hik}^{PS} = M_\gamma \quad (s_\gamma\text{=사후층 표본})$$

  - **최종가중치 $pw_{1,hik}$: 중앙값 기준 사분위 범위 k-배수 특이치 절사**

A. 참고: 표본가중치 상세논의

- **패널가중치**
  - **패널응답성향점수(response propensity score)**

$$\hat{\rho}_t(x_{hik}) = Pr(t\text{차조사 패널응답}|(t-1)\text{차조사 패널응답}, x_{hik})$$

$$= \frac{exp\left(\underline{x}'_{hik}\,\underline{\beta}\right)}{1+exp\left(\underline{x}'_{hik}\,\underline{\beta}\right)} \qquad \& \quad \underline{x}_{hik} = \text{개인 } (hik) \text{ 특성 정보}$$

  - t 차 **추적조사 무응답 조정 가중치**

$$pw_{t,hik} = \begin{cases} \frac{pw^F_{t-1,hik}}{\hat{\rho}_t(x_{hik})} & \text{패널응답자} \\ 0 & \text{패널무응답} \end{cases}$$