분석 패키지 소개 : SPSS

복합표본조사자료분석 : SPSS

서론

- **일반적으로 모평균, 모비율, 모총합 등 모수 추정에 타당하고 신뢰할 만한 결과 제공을 위해 표본조사설계 수립해 수행**
- **표본조사로 수집된 변수들 사이의 내적관계 평가 및 분석에 대한 관심도 증가하는 추세**
- **표준통계패키지는 표본조사의 다양한 설계 특성을 고려하지 않은 분석 방법을 제공함에 주의**
  - 표본추출이론과 추출과정, 가중치 등을 반영해 추정 가능한 통계 패키지의 사용이 필요
  - SUDAAN, STATA, SPSS CS, SAS survey-procedure, R 등

- **표본조사자료분석에서 설계기반분석과 모형기반분석의 의미**
  - 설계기반분석 : 표본설계의 특성을 반영한 분석
    - 층
    - 군집
    - 표본추출율 (혹은 FPC)
    - 표본추출가중치
    - 사후층화 등
  - 모형기반분석
    - 표본설계특성을 고려하지 않고 가정된 모형에 근거한 분석
      - 가정된 모형을 기반으로 분석하므로 추정 및 분석 결과는 가정된 모형에 의존
    - 표본조사에서 무한모집단으로부터 추출된 독립적인 확률추출을 가정한 분석(고전적인 가정)

[참고] 모수 추정은 설계 특성을 반영해 추정해야 하고, 모형 분석은 많은 연구자들의 논쟁이 있지만 일반적으로 표본조사의 설계 특성을 무시한 분석은 편향되고 부정확한 결과 제공 가능성이 높으므로 조사 설계 특성을 고려한 분석이 필요

1. 설계기반분석 과정

- 설계기반분석 수행을 위해 요구되는 단계별 과정과 내용

① 표본설계의 주요한 특성 정보/변수 선정
  - 층 변수, 군집 변수, 추출단계별 추출률 등
  - 유한모집단 수정계수(fpc) 계산을 위한 부모집단 크기 정보 등

② 표본단위에 대한 추출 가중치 계산

③ 무응답, 사후층화 조정 등을 고려해 부여한 조정 가중치 및 최종가중치

④ 설계기반분석을 위한 데이터 파일 구축
  - 개별 표본단위별 층, 집락, 모집단 크기 정보를 연계한 자료 생성
  - 복합표본조사자료분석에서 요구되는 추가 정보 입력

⑤ 분석 방법 결정 및 프로그램 작성

⑥ 분석 수행 및 해석

예제 : 설계기반자료분석

- 표본조사 목적
  - 3개 지역내 사설요양원에서 2007년에 입거한 환자 중 메디케어로 비용을 지불하는 환자 비율 추정
- 표본설계
  - 모집단 : 3개 지역, 43개 사설요양원에 거주하는 환자
    - 모수 추정을 위한 추가 요구 정보
      - 2007년에 입거한 전체 환자 수

지역	사설요양원수	2007년 입거 환자수
1	12	1770
2	20	1200
3	11	5200
계	43	8170

  - 표본추출방법 : 2단계 층화집락추출법
    - 층화 기준 : 지역(region)
    - 집락 : 사설요양원(nurshome)
  - 표본추출과정
    - 1단계(psu=사설요양원) : 각 층마다 3개의 사설 요양원을 추출
    - 2단계(ssu=환자) : 표본 사설요양원마다 10명의 환자를 추출

- 설계기반분석의 단계별 과정
  - ① 표본설계의 주요한 특성 정보/변수 선정
    - 층화 기준 : 지역(region)
    - 집락 : 사설요양원(nurshome)
    - 추출과정
      - 1차 추출(psu) : 사설요양원(WOR)
      - 2차 추출(ssu) : 환자
    - 추출률 계산에 필요한 정보
      - psu : 층(지역)내 사설요양원수(nrgnhomes : Mh)
      - ssu : 사설요양원내 2007년 입거한 환자 수(nhadmiss)
  - ② 기본(설계) 가중치 계산
    - 1차 및 2차 단위 추출률 $f_{1h} = \frac{3}{M_h}, \quad f_{2hi} = \frac{10}{N_{hi}}$
    - 개별 표본의 최종 추출률 $f_{hi} = f_{1h} f_{2hi} = \frac{3}{M_h} \frac{10}{N_{hi}} = \frac{30}{M_h N_{hi}}$
    - 표본 가중치 $w_{hi} = \frac{M_h N_{hi}}{30}$

③ 무응답 및 사후층화 보정을 반영한 최종 가중치

- 무응답 조정 가중치 : 완전 응답으로 무응답 조정 가중치 미부여
- 사후층화 조정 가중치 : 환자들의 연령, 성별 등에 의한 사후 조정 계획은 없으므로 사후층화 보정 가중치 미부여

$$w_t = w_{hi} = \frac{M_h N_{hi}}{30}$$

  - 2007년에 입거한 환자수에 대한 사후층화 조정 검토 : 기본 가중치를 부여한 환자수가 층별로 차이가 있으므로 이를 보정하기 위해 층별 환자수를 기준으로 사후층화 조정

$$w_f = w_t \frac{N_h}{\sum w_{hi}}$$

④ 설계기반분석 자료 구축

- 엑셀 자료 참조

- 조사 자료

region	nurshome	patient	medicaid	nrgnhomes	nhadmiss	nhad_pop	f_1	f_2	wt
1	1	1	1	12	123	1770	0.25000	0.08130	49.200
1	1	2	1	12	123	1770	0.25000	0.08130	49.200
1	1	3	0	12	123	1770	0.25000	0.08130	49.200
1	1	4	0	12	123	1770	0.25000	0.08130	49.200
1	1	5	0	12	123	1770	0.25000	0.08130	49.200
1	1	6	0	12	123	1770	0.25000	0.08130	49.200
1	1	7	0	12	123	1770	0.25000	0.08130	49.200
1	1	8	0	12	123	1770	0.25000	0.08130	49.200
1	1	9	0	12	123	1770	0.25000	0.08130	49.200
1	1	10	0	12	123	1770	0.25000	0.08130	49.200
1	2	1	1	12	89	1770	0.25000	0.11236	35.600
1	2	2	0	12	89	1770	0.25000	0.11236	35.600
1	2	3	0	12	89	1770	0.25000	0.11236	35.600
1	2	4	0	12	89	1770	0.25000	0.11236	35.600
1	2	5	0	12	89	1770	0.25000	0.11236	35.600
1	2	6	0	12	89	1770	0.25000	0.11236	35.600
1	2	7	0	12	89	1770	0.25000	0.11236	35.600
1	2	8	0	12	89	1770	0.25000	0.11236	35.600
1	2	9	0	12	89	1770	0.25000	0.11236	35.600

⑤ 복합표본조사자료분석

- 전문 통계패키지 선택 및 프로그램 작성 : SPSS 사용
  - SUDAAN, STATA, SPSS CS, SAS survey-procedure, R 등 가능

⑥ 분석 실행 및 결과 해석

- 사후층화 조정 가중치 부여 전 결과

층	모집단	가중합	사용	미사용	사후 조정비
1	1770	1772.00	503.60	1268.40	0.99887
2	1200	986.67	568.00	418.67	1.21622
3	5200	4315.67	3320.53	995.13	1.20491
전체	8170	7074.33	4392.13	2682.20	1.15488

복합 표본: 기술통계

일변량 통계량

		추정값	표준오차
평균	메디카드사용여부	.62	.053
합계	메디카드사용여부	4392	546.784

일변량 통계량

지역_층			추정값	표준오차
1	평균	메디카드사용여부	.28	.092
	합계	메디카드사용여부	504	278.583
2	평균	메디카드사용여부	.58	.073
	합계	메디카드사용여부	568	105.768
3	평균	메디카드사용여부	.77	.080
	합계	메디카드사용여부	3321	458.451

– 사후층화 조정 가중치 반영 후 결과

- 사후층화 조정 비

$$w_{ps,h} = \frac{N_h}{\hat{N}_h} = \frac{N_h}{\sum_h wt_{h1}}$$

- 최종 가중치

$$wt_{h2} = w_{ps,h} wt_{h1}$$

일변량 통계량

		추정값	표준오차
평균	메디카드사용여부	.64	.054
합계	메디카드사용여부	5195	631.759

일변량 통계량

지역_층			추정값	표준오차
1	평균	메디카드사용여부	.28	.092
	합계	메디카드사용여부	503	278.269
2	평균	메디카드사용여부	.58	.073
	합계	메디카드사용여부	691	128.636
3	평균	메디카드사용여부	.77	.080
	합계	메디카드사용여부	4001	552.393

SPSS 복합표본

SPSS복합표본 : complex samples wizards

- 표본설계/추출
  - sampling wizard : sampling plan
    - sample selection with random sampling methods
    - give a proper estimation methods

- 표본조사자료분석
  - analysis preparation wizard : analysis plan
    - complex samples plan(CS plan) : include sample structure, the estimation method per each stage, sampling weights
    - variance estimation

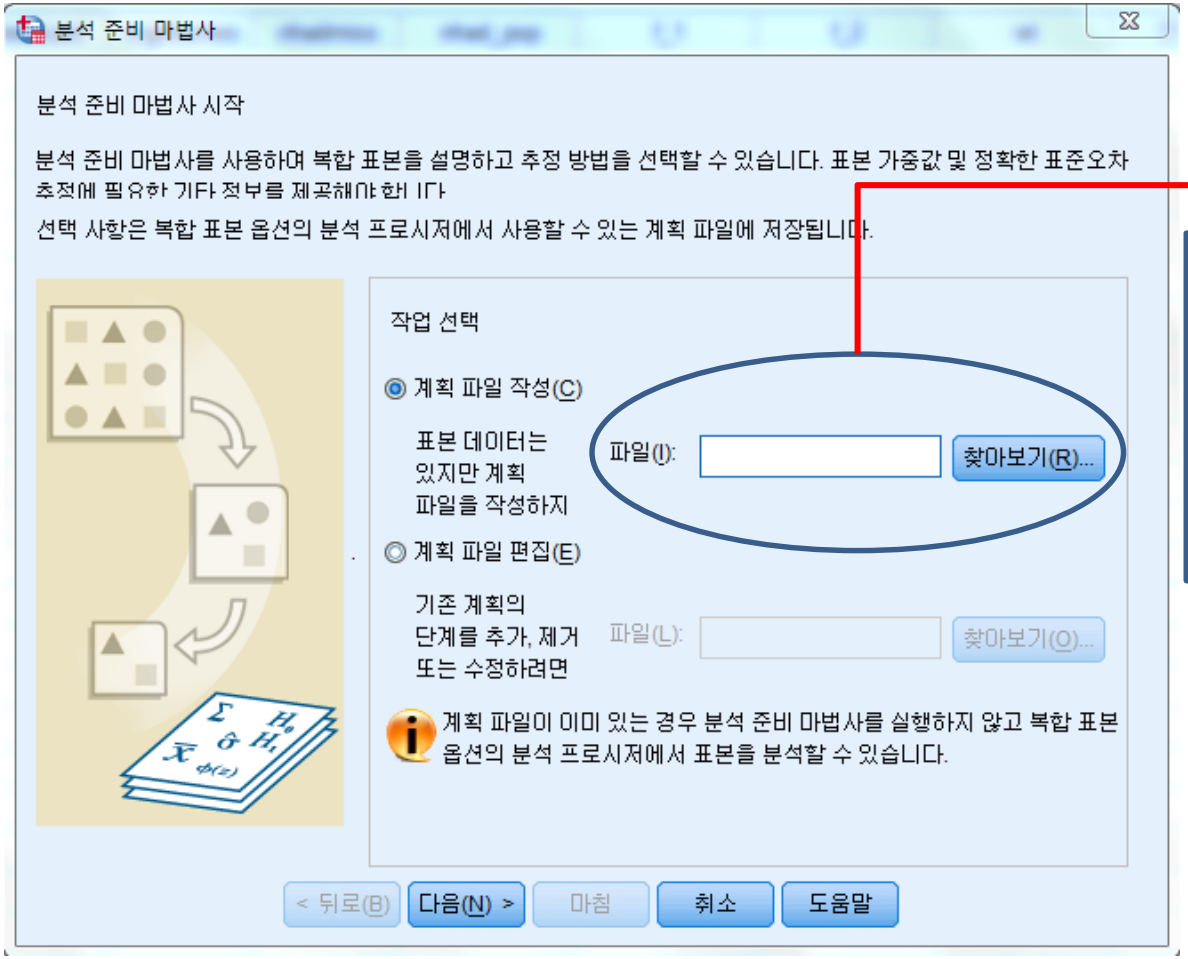
- **SPSS 복합표본조사 자료분석**

SPSS 복합표본

1. **표본추출 기능 :**
   - **모집단 입력**
   - **확률추출**
2. **복합표본 분석**
   - **1단계(분석준비)**
     : 분석자료의 표본설계 입력
   - **2단계(분석)**
     : 빈도분석/기술통계
     : 회귀분석/분산분석
     : 로지스틱회귀분석 등

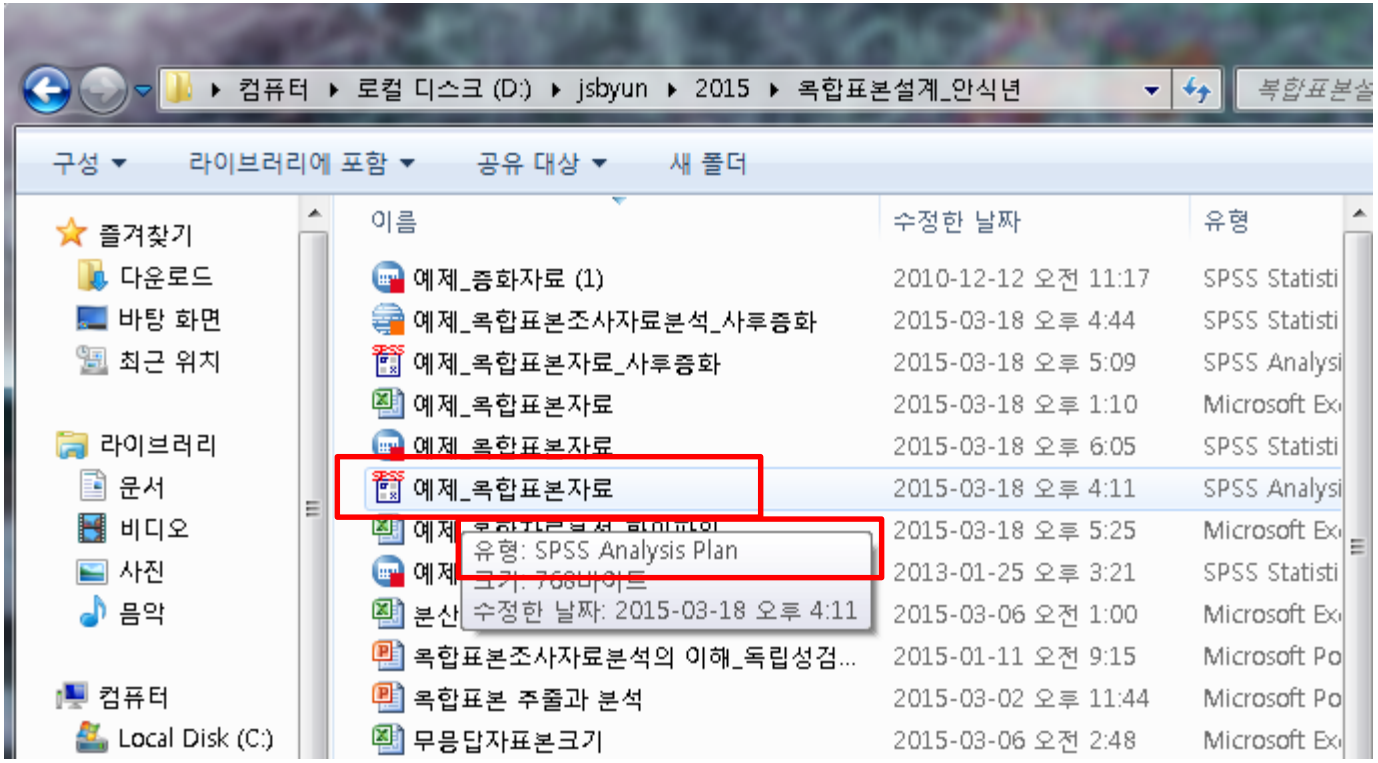
SPSS복합표본 조사자료분석

- **표본설계 과정 요약**



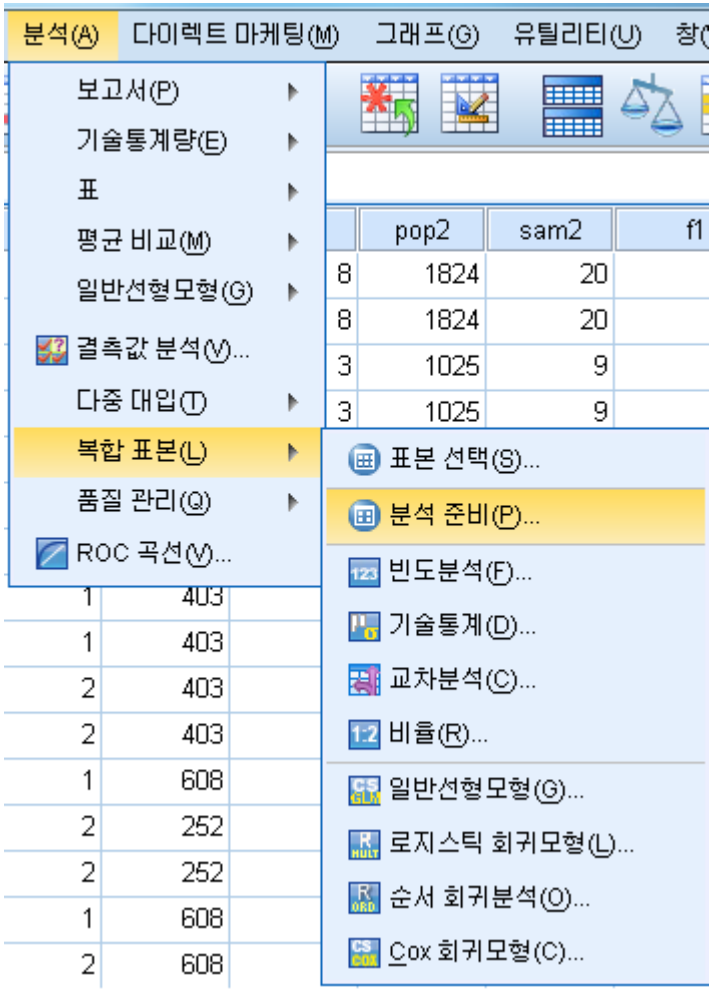
- 표본자료만 있으므로 "계획파일 작성"을 선택
- 계획 파일을 저장할 경로를 지정하기 위해 찾아보기를 눌러 "계획파일명을 입력"한 후 저장해 지정

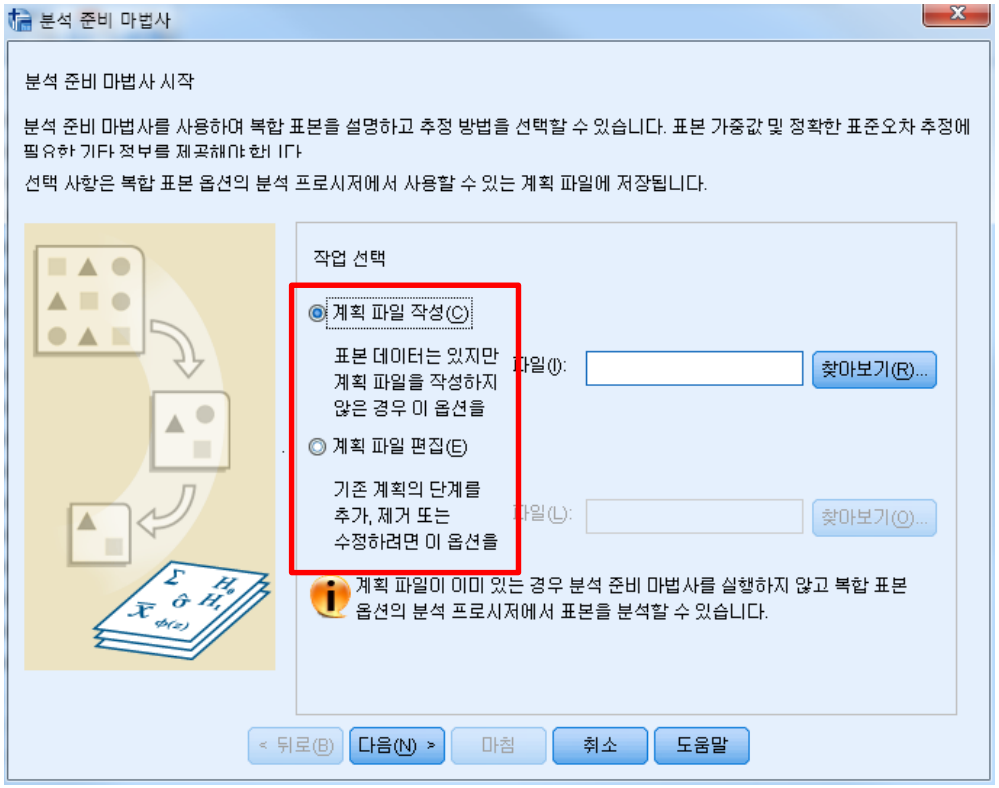
- **계획파일 확인**


컴퓨터 ▸ 로컬 디스크 (D:) ▸ jsbyun ▸ 2015 ▸ 복합표본설계_안식년
구성
라이브러리에 포함
공유 대상
새 폴더
즐겨찾기
다운로드
바탕 화면
최근 위치
라이브러리
문서
비디오
사진
음악
컴퓨터
Local Disk (C:)
이름
수정한 날짜
유형
예제_층화자료 (1) 2010-12-12 오전 11:17 SPSS Statisti
예제_복합표본조사자료분석_사후층화 2015-03-18 오후 4:44 SPSS Statisti
예제_복합표본자료_사후층화 2015-03-18 오후 5:09 SPSS Analysi
예제_복합표본자료 2015-03-18 오후 1:10 Microsoft Ex
2015-03-18 오후 6:05 SPSS Statisti
예제_복합표본자료 2015-03-18 오후 4:11 SPSS Analysi
2015-03-18 오후 5:25 Microsoft Ex
유형: SPSS Analysis Plan
수정한 날짜: 2015-03-18 오후 4:11
2013-01-25 오후 3:21 SPSS Statisti
2015-03-06 오전 1:00 Microsoft Ex
복합표본조사자료분석의 이해_독립성검... 2015-01-11 오전 9:15 Microsoft Po
복합표본 주줄과 분석 2015-03-02 오후 11:44 Microsoft Po
무응답자표본크기 2015-03-06 오전 2:48 Microsoft Ex

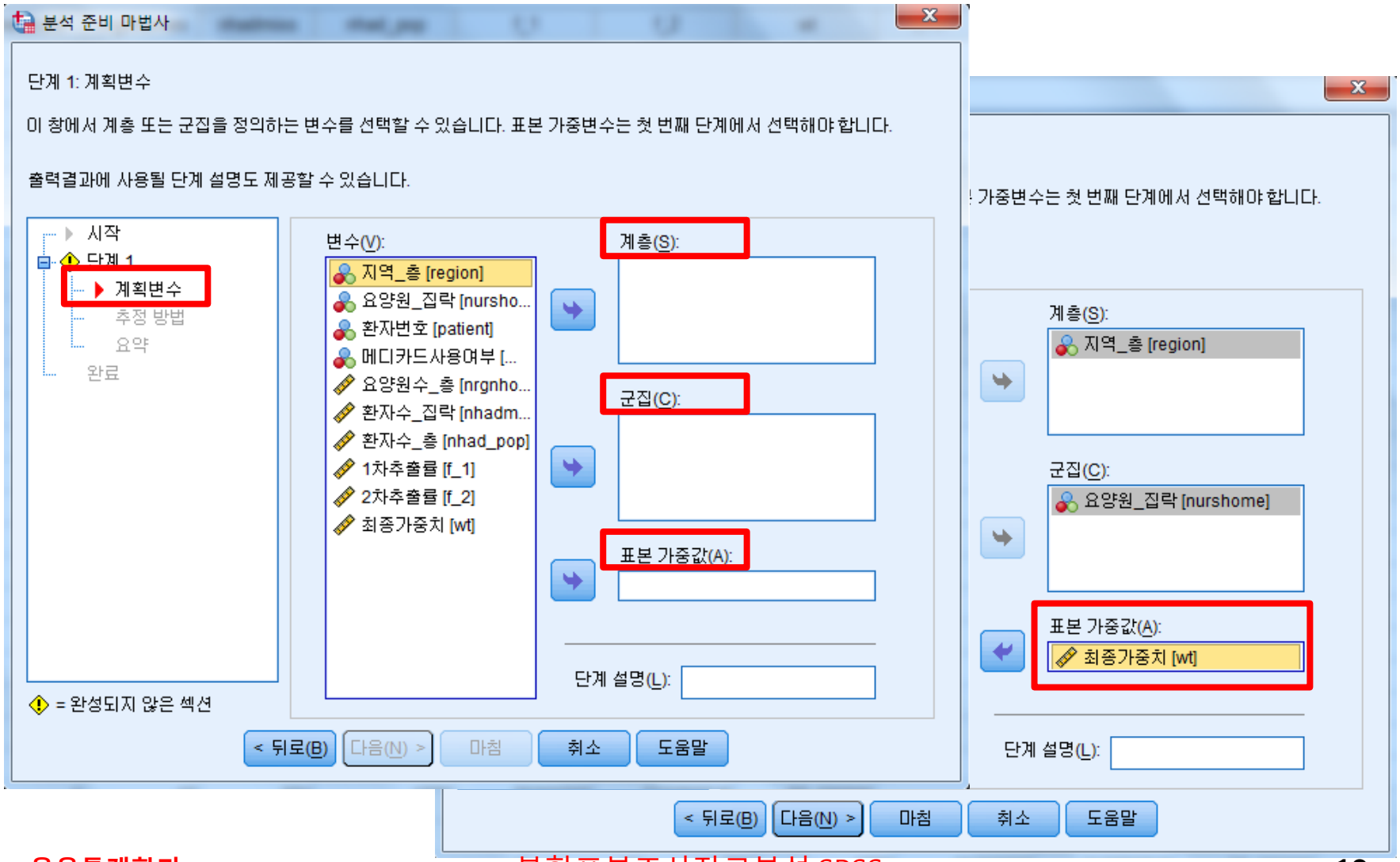
- **1단계 : 복합표본설계 내용 입력**

  - 층/집락 변수 설정
  - 최종 가중값 변수 설정
  - 추정방법 설정 : WR, WOR
  - 모집단 정보 설정 : 모집단 크기 혹은 추출확률

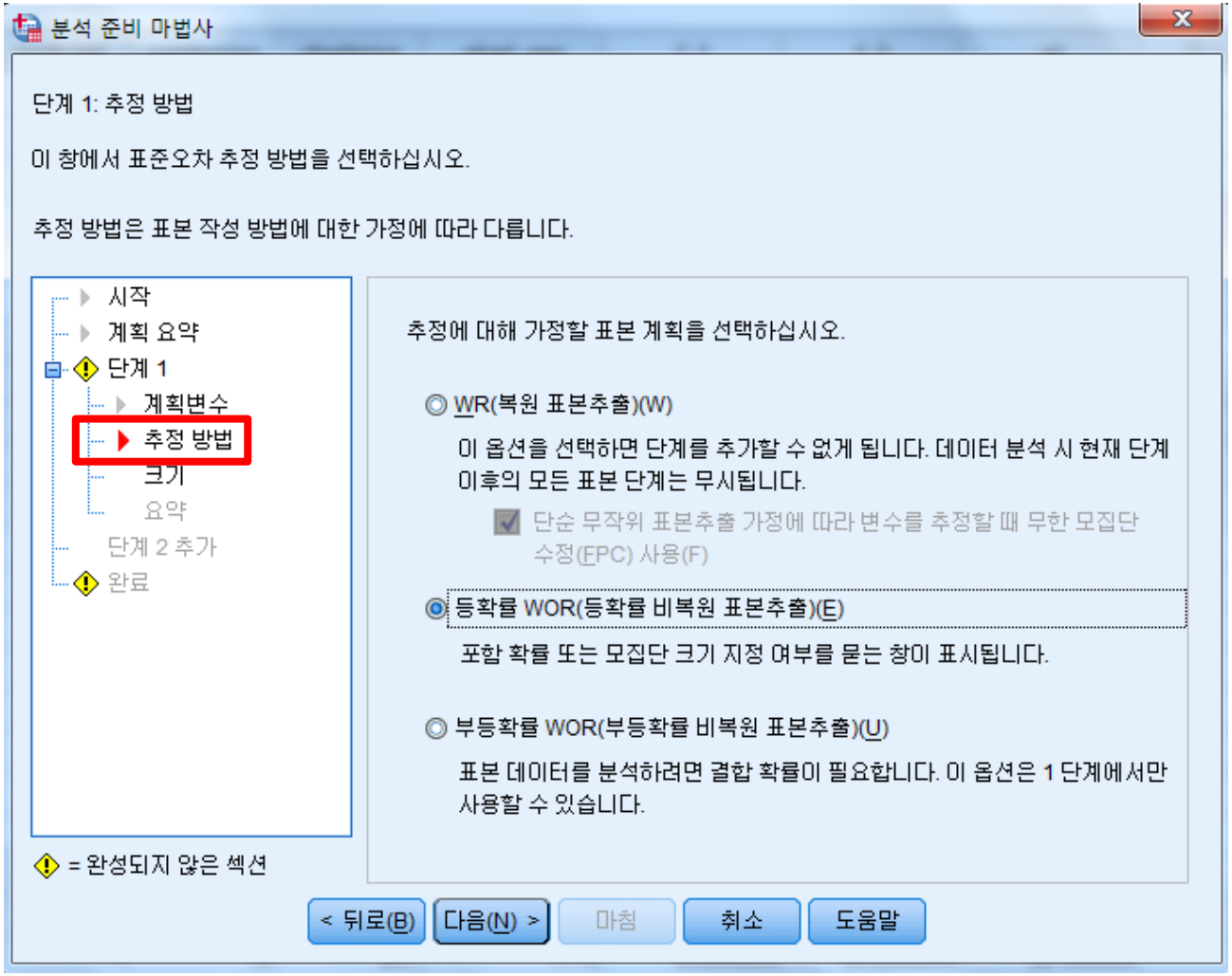




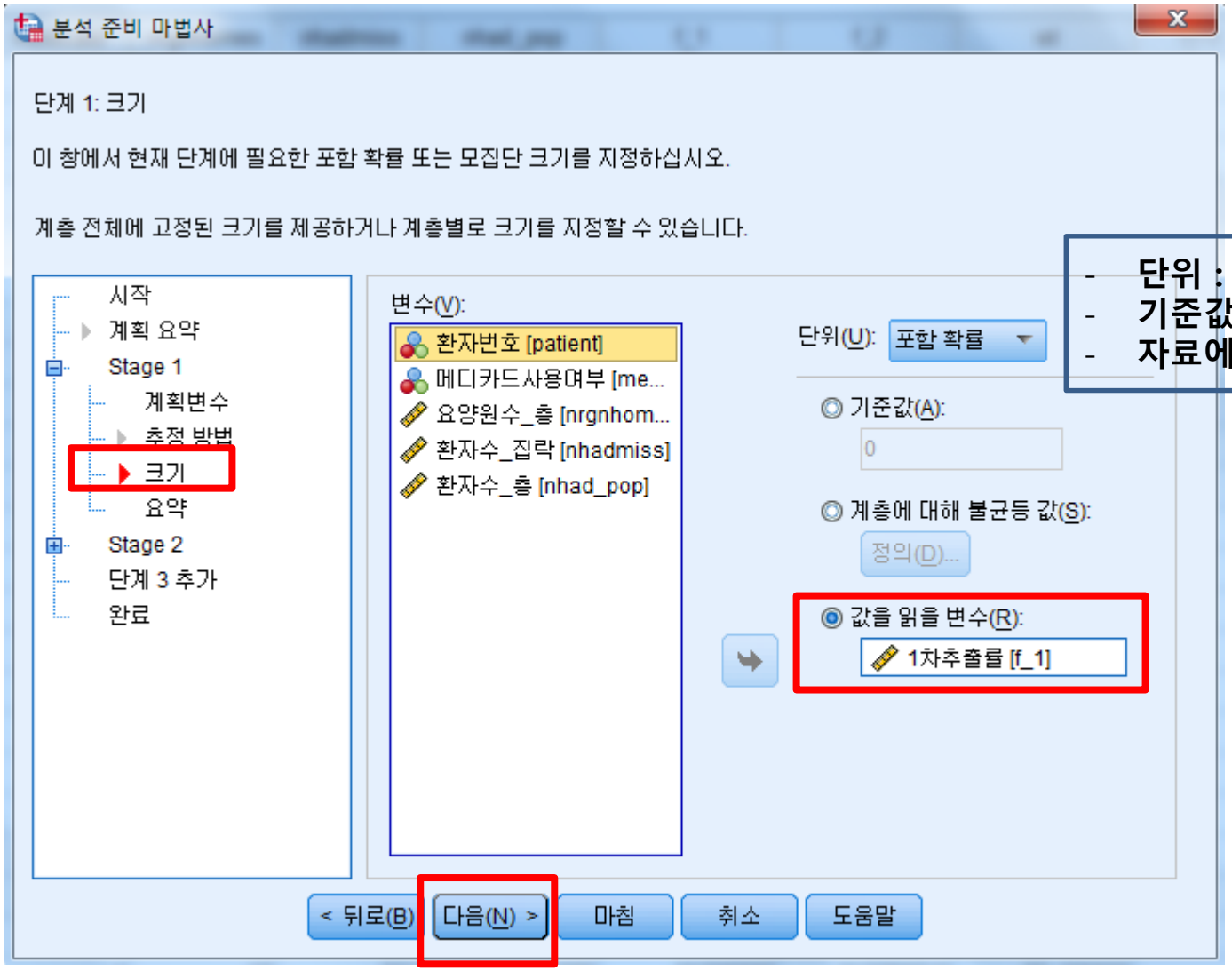
- **표본설계 변수 정의**
  - 분석할 표본조사 자료의 표본설계 내용을 정의하는 단계



- **표준오차 추정 방법 선택**
  - **단계 1 : psu에 대한 표준오차 추정 방법 설정**

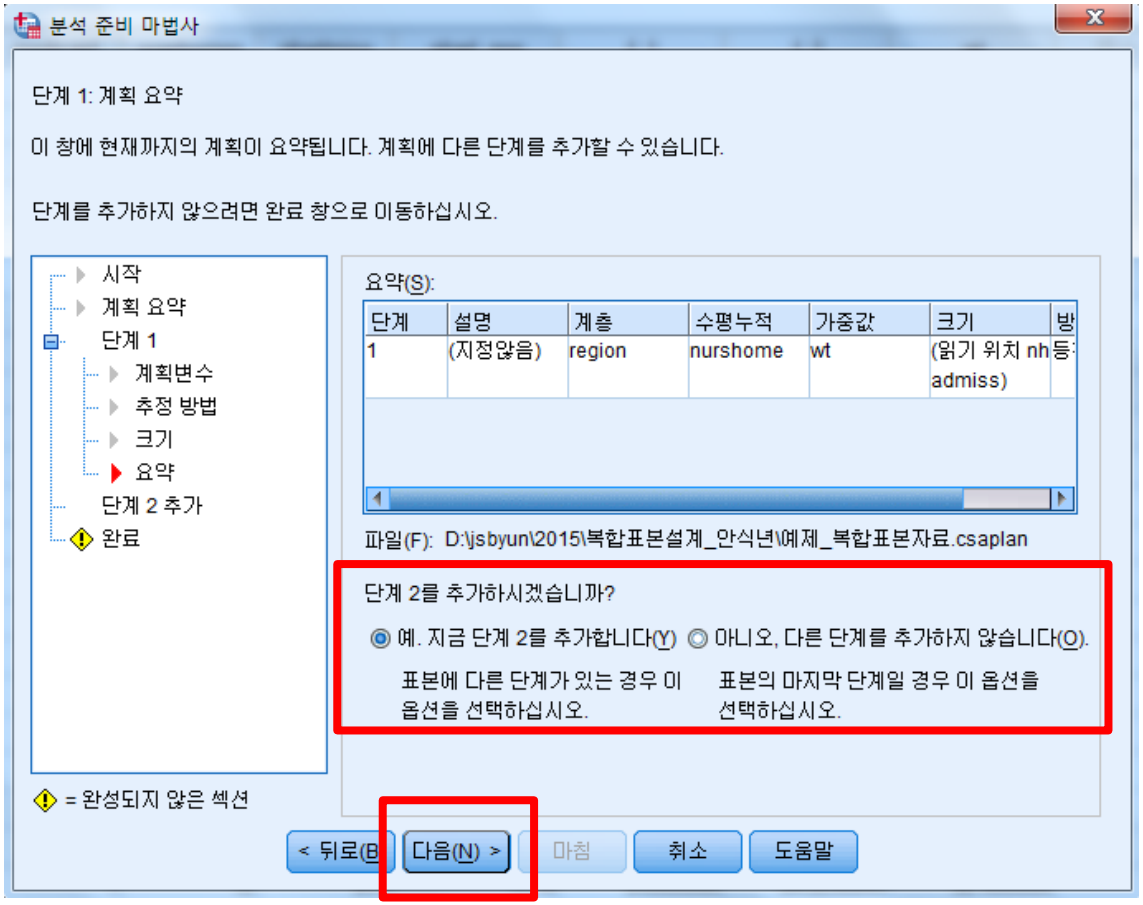


- **부모집단 크기 정의 : 1차(psu) 추출률**

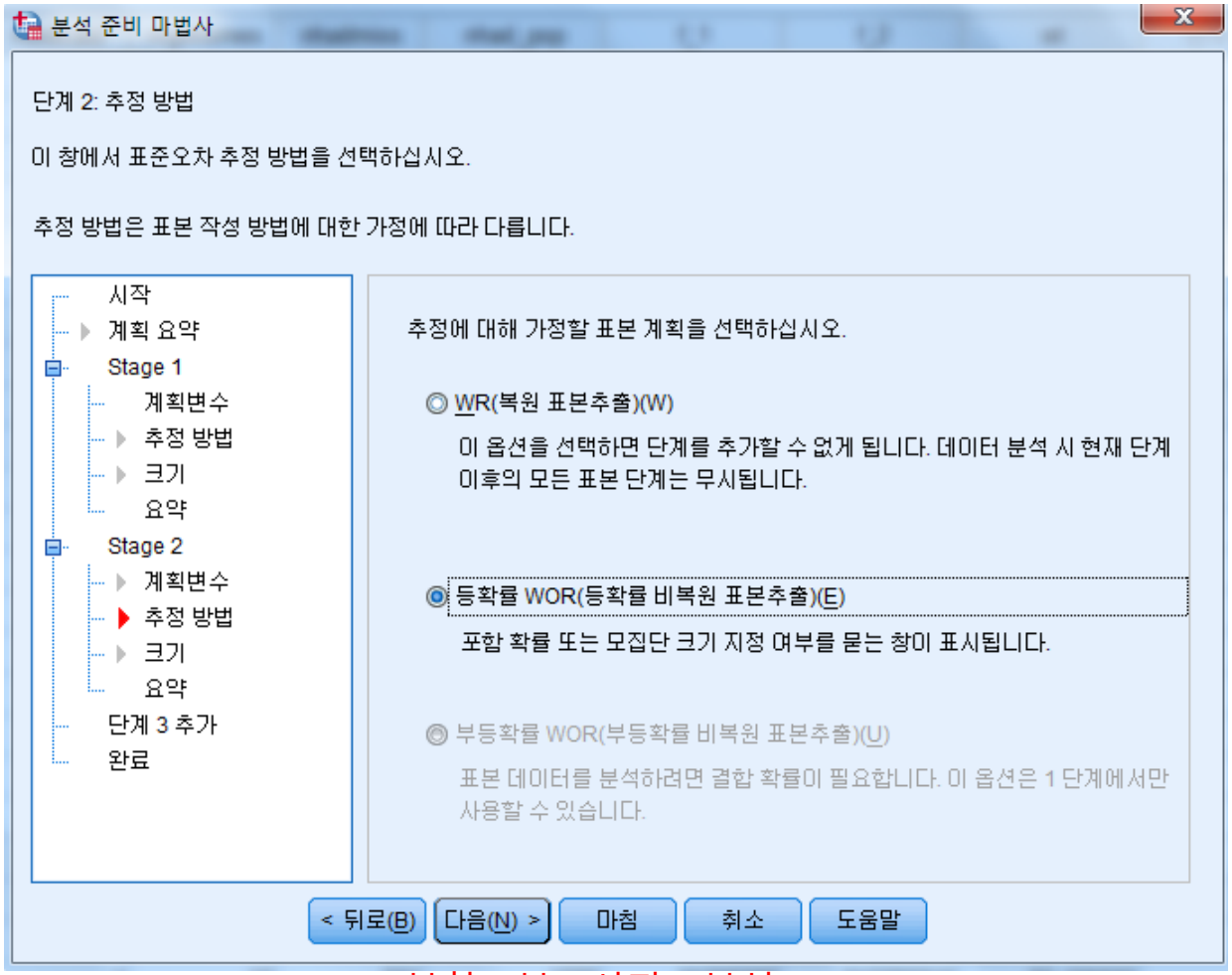


- 단위 : 포함확률/모집단크기
- 기준값 선언 가능
- 자료에서 변수로 선언가능

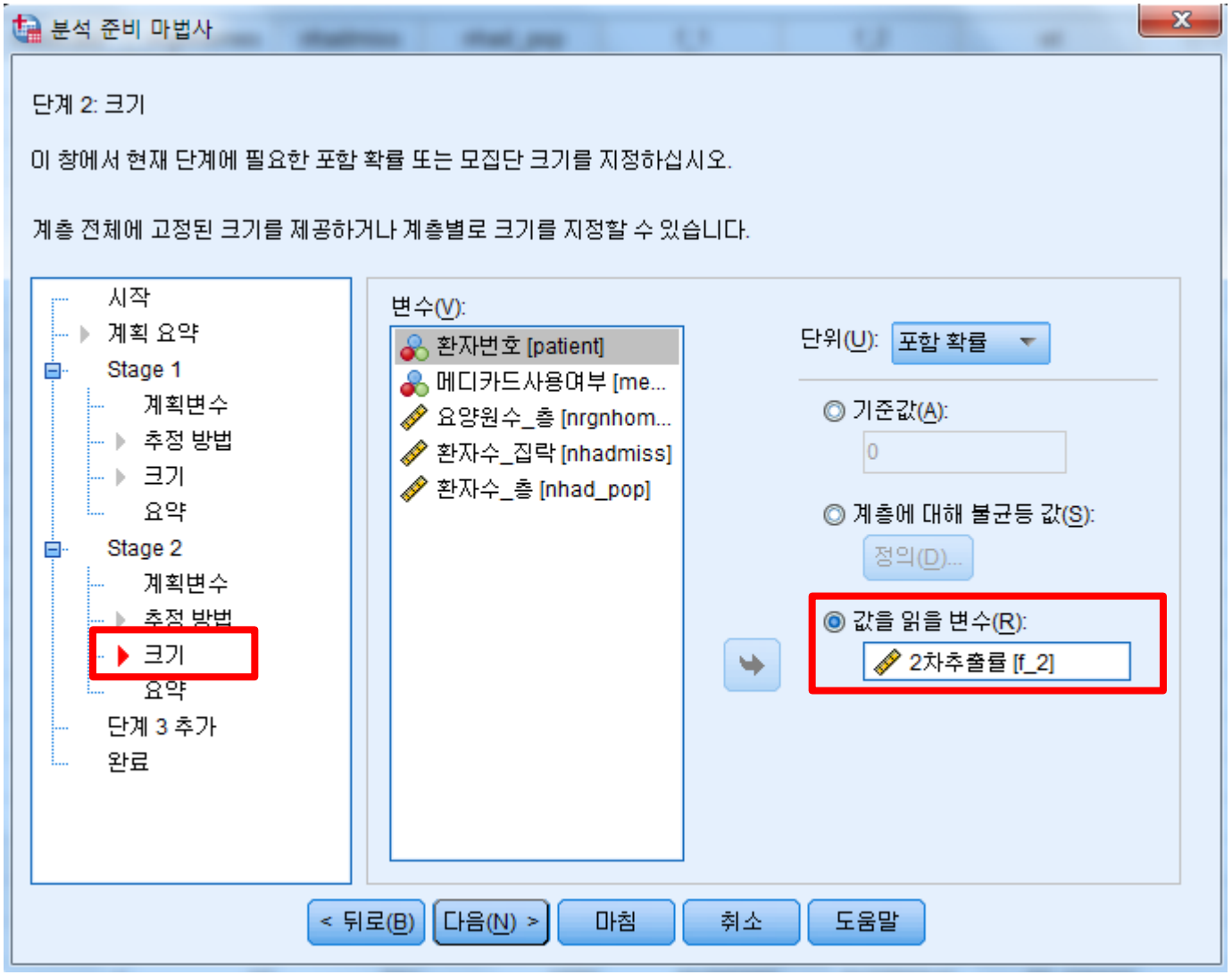
- **다단계 추출단위의 추가**
  - **2차 이상의 추출단위 지정하는 단계**



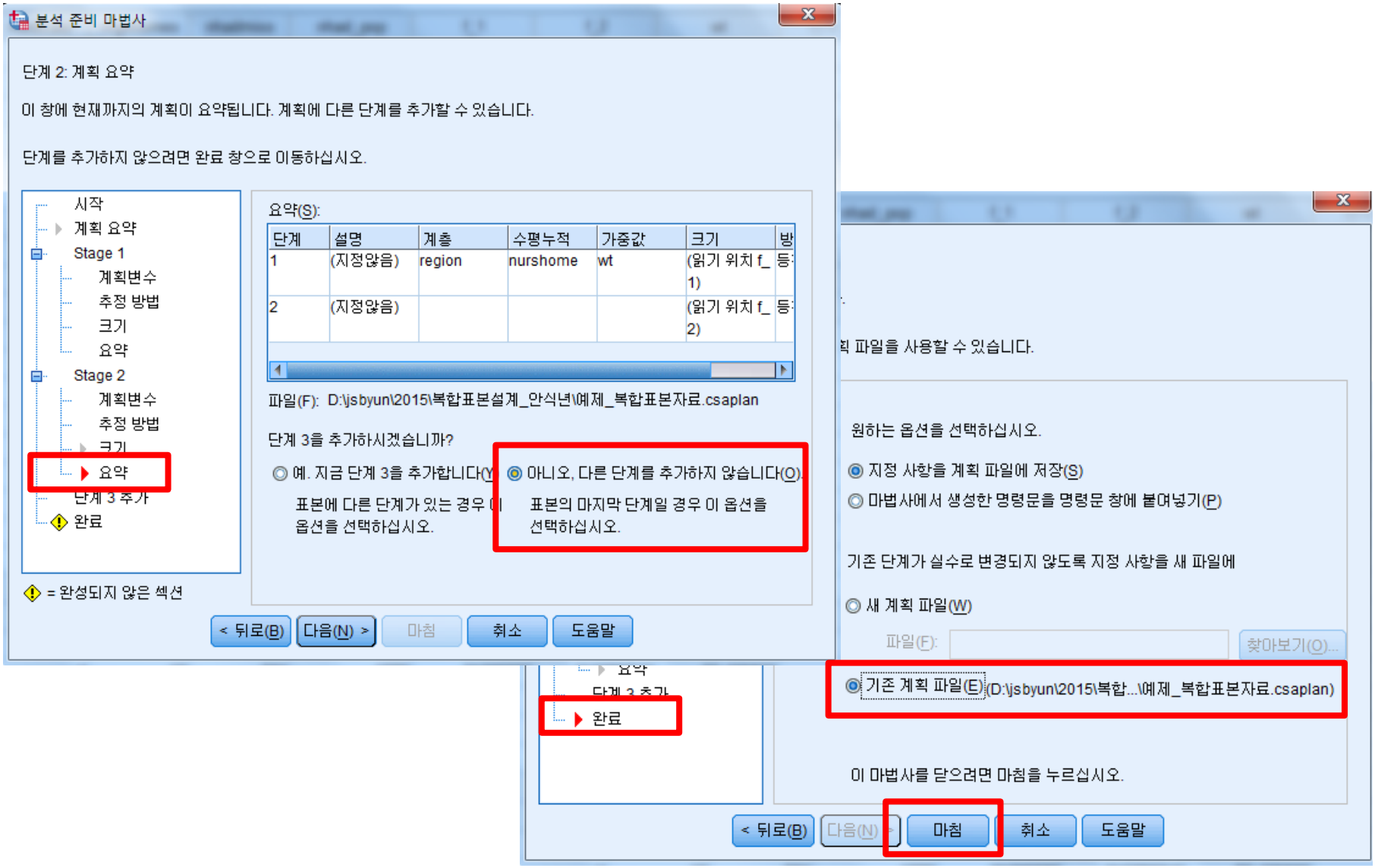
- **2차 추출단위 추정 방법 선택(psu와 동일)**
  - **계획변수는 동일하므로 추가 설정하지 않음**



- **부모집단 크기 정의 : 2차(ssu) 추출률**



- **분석 자료에 대한 표본설계 과정 요약 및 저장**



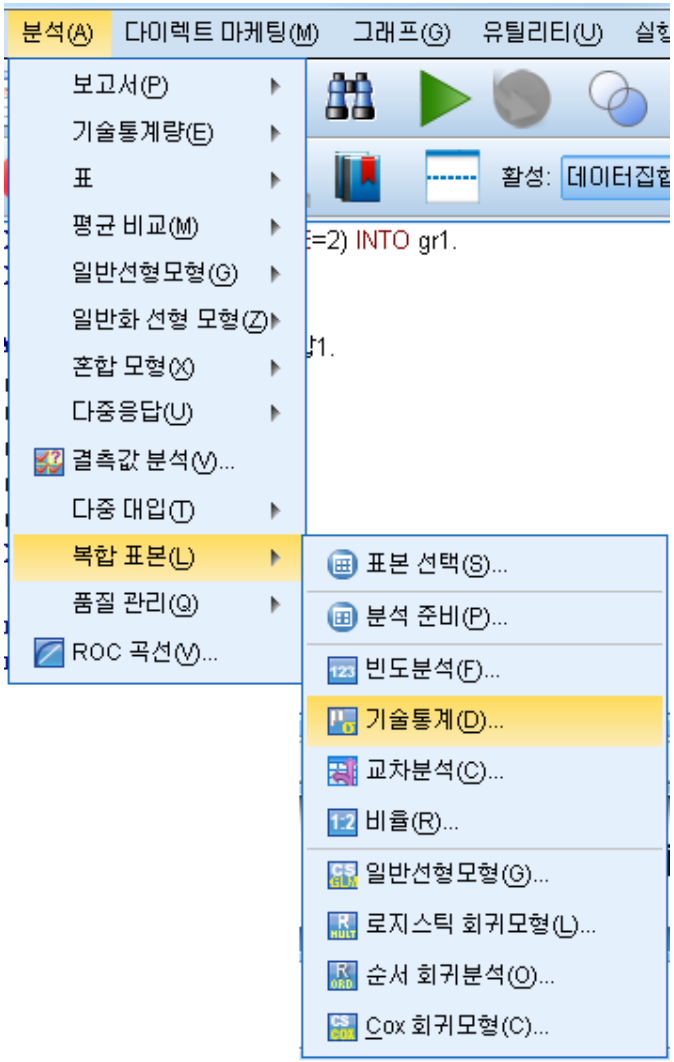
- 복합표본조사 자료분석을 위한 준비 과정 요약 결과(출력창 제공)

요약

			단계 1	단계 2
계획변수	층화	1	지역_층	
	군집	1	요양원_집락	
분석 정보	추정량 가정		비복원 등확률 표본추출	비복원 등확률 표본추출
	포함 확률		변수 1차추출률에서 확보	변수 2차추출률에서 확보

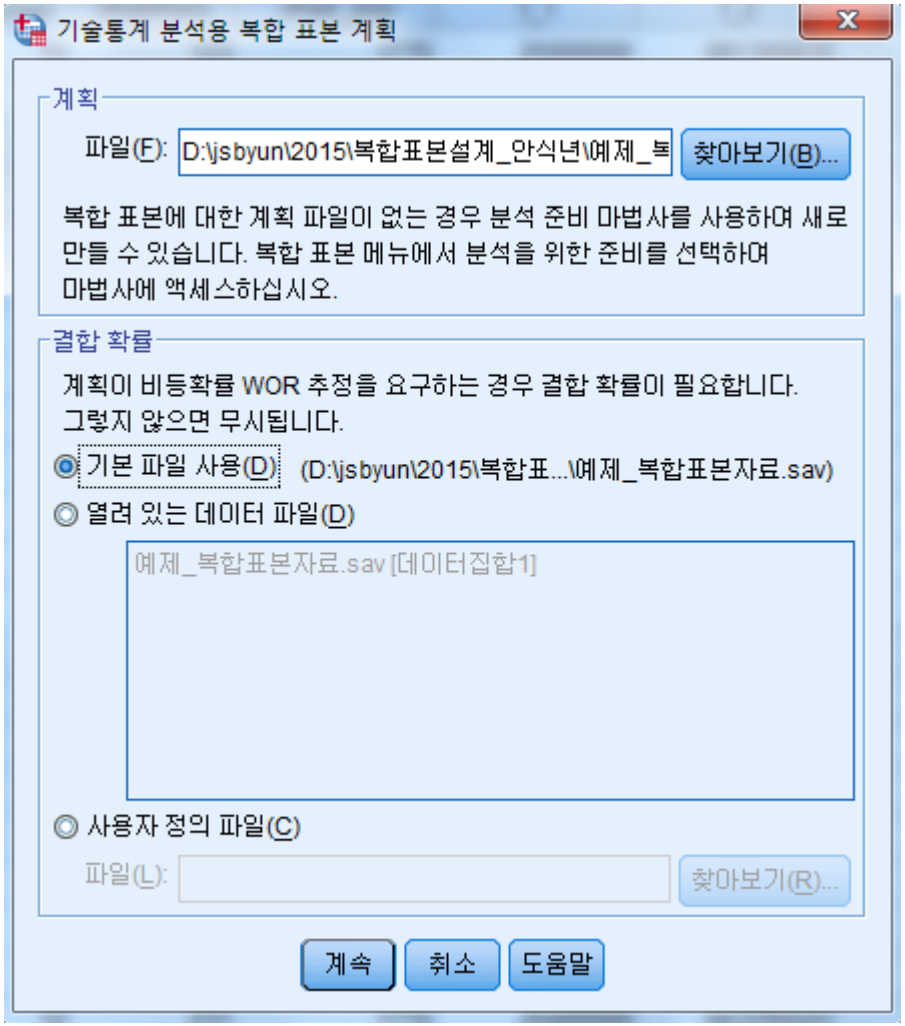
계획 파일: D:\jsbyun\2015\복합표본설계_안식년\예제_복합표본자료.csaplan
가중변수: 최종가중치
SRS 추정량: 비복원 표본추출

- **2단계 : 복합표본조사 자료 분석**

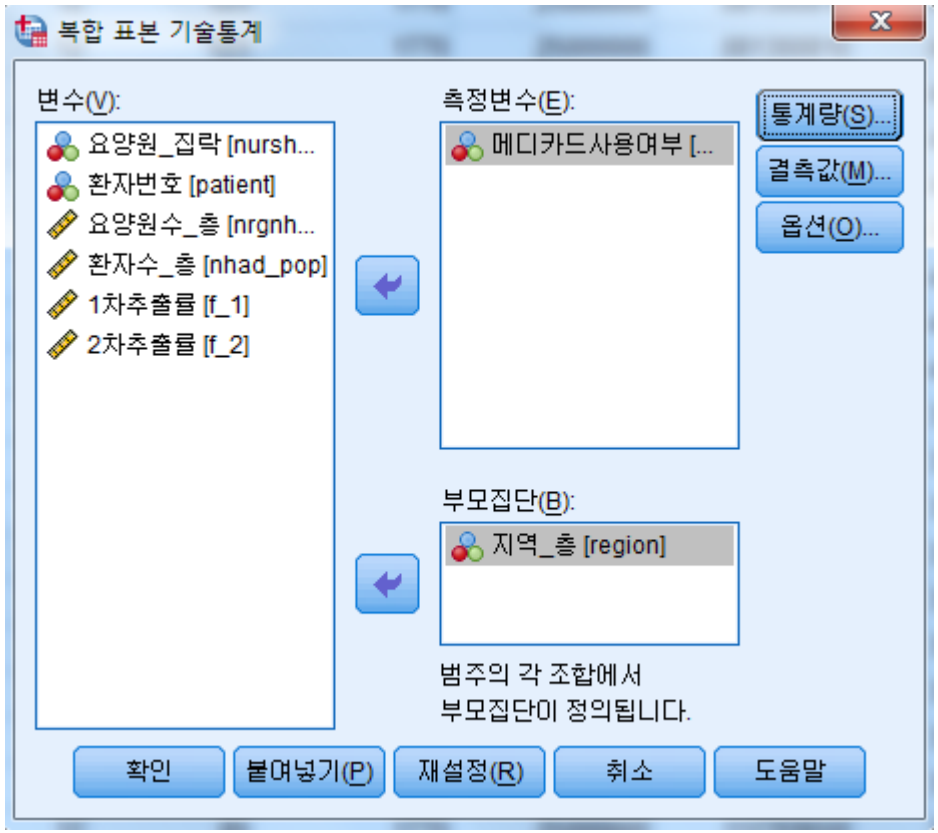


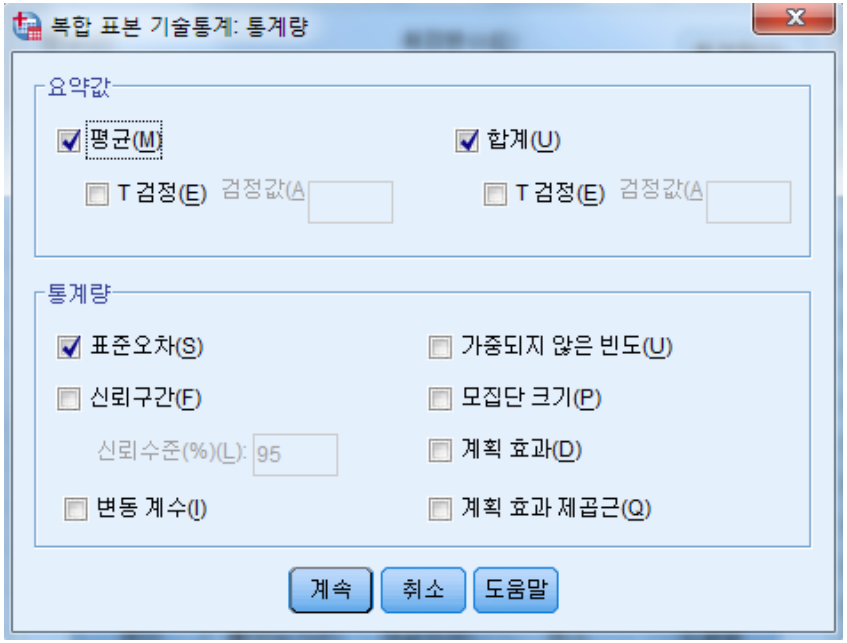
- 빈도분석/기술통계
- 교차분석(카이제곱검정)/비율 추정
- 일반선형모형/로지스틱 회귀모형
- 순서형 회귀분석/생존분석

[1]복합표본조사 자료분석 : 기술통계



- **기술통계 분석 단위 및 통계량 설정**





- **출력 결과 : 복합표본자료분석**

복합 표본: 기술통계

일변량 통계량

		추정값	표준오차
평균	메디카드사용여부	.62	.053
합계	메디카드사용여부	4392	546.784

부모집단 기술통계

일변량 통계량

지역_층			추정값	표준오차
1	평균	메디카드사용여부	.28	.092
	합계	메디카드사용여부	504	278.583
2	평균	메디카드사용여부	.58	.073
	합계	메디카드사용여부	568	105.768
3	평균	메디카드사용여부	.77	.080
	합계	메디카드사용여부	3321	458.451

- **참고 : 가중치를 부여한 분석**
  - **복합표본자료분석 미적용**

비가중

보고서

메디카드사용여부

지역_층	평균	N	표준편차	평균의 표준오차
1	.23	30	.430	.079
2	.57	30	.504	.092
3	.77	30	.430	.079
합계	.52	90	.502	.053

$$se = sd/\sqrt{m}$$

가중치 부여

보고서

메디카드사용여부

지역_층	평균	N	표준편차	평균의 표준오차
1	.28	1772	.451	.011
2	.58	987	.494	.016
3	.77	4316	.421	.006
합계	.62	7074	.485	.006

- **출력 결과 : 일반분석과 복합표본자료분석 결과의 비교**

복합 표본: 기술통계

일변량 통계량

		추정값	표준오차
평균	메디카드사용여부	.62	.053
합계	메디카드사용여부	4392	546.784

부모집단 기술통계

일변량 통계량

지역_층			추정값	표준오차
1	평균	메디카드사용여부	.28	.092
	합계	메디카드사용여부	504	278.583
2	평균	메디카드사용여부	.58	.073
	합계	메디카드사용여부	568	105.768
3	평균	메디카드사용여부	.77	.080
	합계	메디카드사용여부	3321	458.451

비가중

보고서

메디카드사용여부

지역_층	평균	N	표준편차	평균의 표준오차
1	.23	30	.430	.079
2	.57	30	.504	.092
3	.77	30	.430	.079
합계	.52	90	.502	.053

모든 추정치 편향

가중치 부여

보고서

메디카드사용여부

지역_층	평균	N	표준편차	평균의 표준오차
1	.28	1772	.451	.011
2	.58	987	.494	.016
3	.77	4316	.421	.006
합계	.62	7074	.485	.006

모수 비편향 SE 편향

추가 : 요양원 입거 노인 환자수 추정/사후층화 과정

- 층별 모집단 크기 검토 및 사후층화 조정 가중치 산출

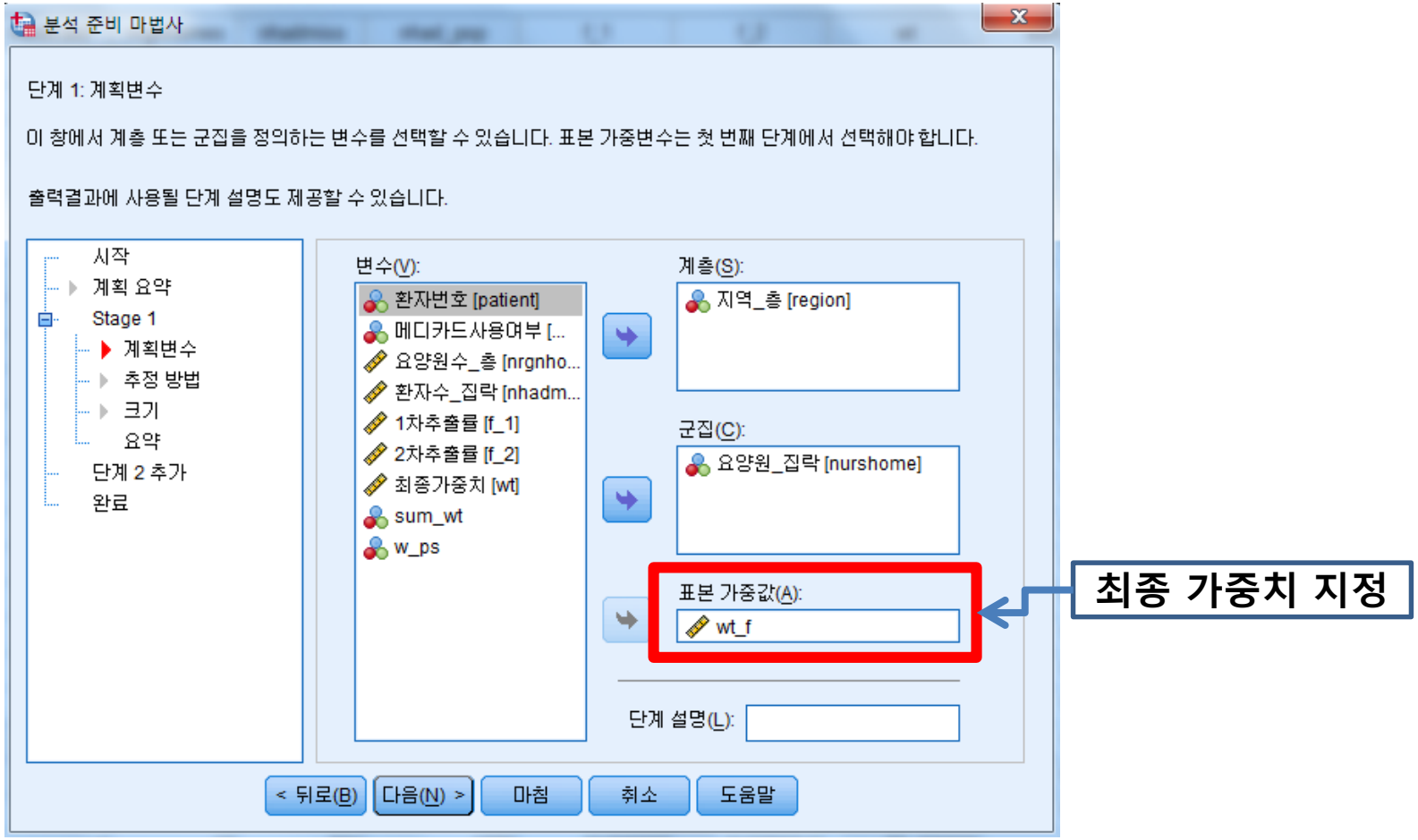
층	모집단	가중합	사용	미사용	사후조정비
1	1770	1772.00	503.60	1268.40	0.99887
2	1200	986.67	568.00	418.67	1.21622
3	5200	4315.67	3320.53	995.13	1.20491
전체	8170	7074.33	4392.13	2682.20	1.15488

- 사후조정 가중치 $$w_{ps,h} = \frac{N_h}{\hat{N}_h} = \frac{N_h}{\sum_h wt_{1,h}}$$

- 최종 가중치 $$wt_{2,h} = w_{ps,h} wt_{1,h}$$

[사후층화 조정 가중치를 반영한 SPSS 복합표본]

- 일반적인 복합표본조사 자료분석과정과 동일하지만 최종 가중치를 사후층화 조정이 반영된 가중치를 선언하여 분석



- 사후층화 조정을 반영한 복합표본조사 자료분석의 분석 준비 과정 요약 결과(SPSS 출력창 제공)

요약

		단계 1	단계 2
계획변수	층화 1	지역_층	
	군집 1	요양원_집락	
분석 정보	추정량 가정	비복원 등확률 표본추출	비복원 등확률 표본추출
	포함 확률	변수 1차추출률에서 확보	변수 2차추출률에서 확보

계획 파일: D:\jsbyun\2015\복합표본설계_안식년\예제_복합표본자료_사후층화.csaplan
가중변수: wt_f
SRS 추정량: 비복원 표본추출

- **사후가중치를 반영한 복합표본자료분석 결과**
  - **사후가중치를 부여하기 전 결과와 약간 차이가 있음**

메디카드사용여부

		추정값	표준오차
모집단 크기	0	2975.206	459.023
	1	5194.793	631.759
	합계	8169.999	592.293

일변량 통계량

		추정값	표준오차
평균	메디카드사용여부	.64	.054
합계	메디카드사용여부	5195	631.759

부모집단 표

메디카드사용여부

지역_층			추정값	표준오차
1	모집단 크기	0	1266.968	197.145
		1	503.032	278.269
		합계	1770.000	444.362
2	모집단 크기	0	509.189	63.483
		1	690.811	128.636
		합계	1200.000	84.242
3	모집단 크기	0	1199.048	409.641
		1	4000.951	552.393
		합계	5200.000	382.435

부모집단 기술통계

일변량 통계량

지역_층			추정값	표준오차
1	평균	메디카드사용여부	.28	.092
	합계	메디카드사용여부	503	278.269
2	평균	메디카드사용여부	.58	.073
	합계	메디카드사용여부	691	128.636
3	평균	메디카드사용여부	.77	.080
	합계	메디카드사용여부	4001	552.393

- **출력 결과 : 일반분석과 복합표본자료분석 결과의 비교**

복합표본자료분석

일변량 통계량

		추정값	표준오차
평균	메디카드사용여부	.64	.054
합계	메디카드사용여부	5195	631.759

부모집단 기술통계

일변량 통계량

지역_층			추정값	표준오차
1	평균	메디카드사용여부	.28	.092
	합계	메디카드사용여부	503	278.269
2	평균	메디카드사용여부	.58	.073
	합계	메디카드사용여부	691	128.636
3	평균	메디카드사용여부	.77	.080
	합계	메디카드사용여부	4001	552.393

가중치 부여/일반 분석

보고서

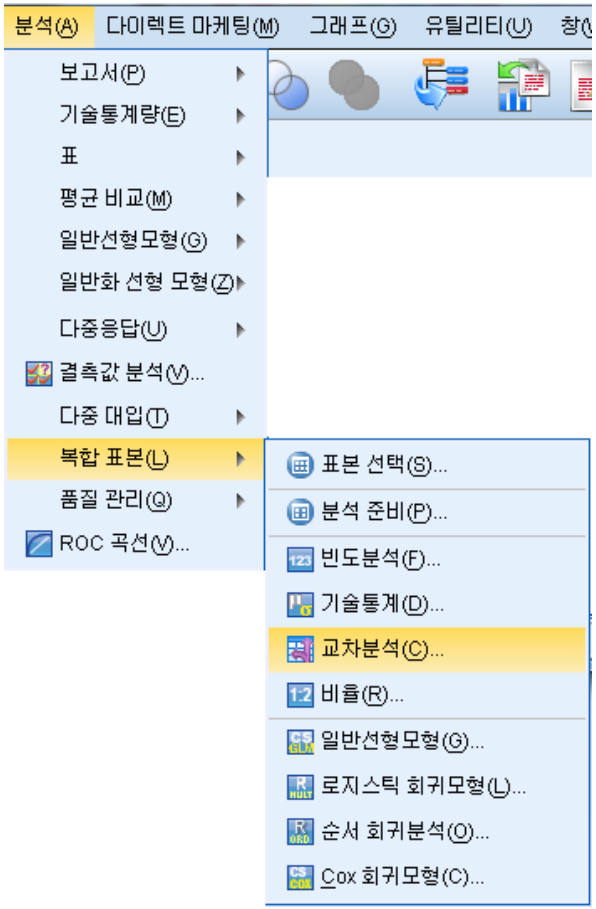
메디카드사용여부

지역_층	합계	평균	N	표준편차	평균의 표준오차
1	503	.28	1770	.451	.011
2	691	.58	1200	.494	.014
3	4001	.77	5200	.421	.006
합계	5195	.64	8170	.481	.005

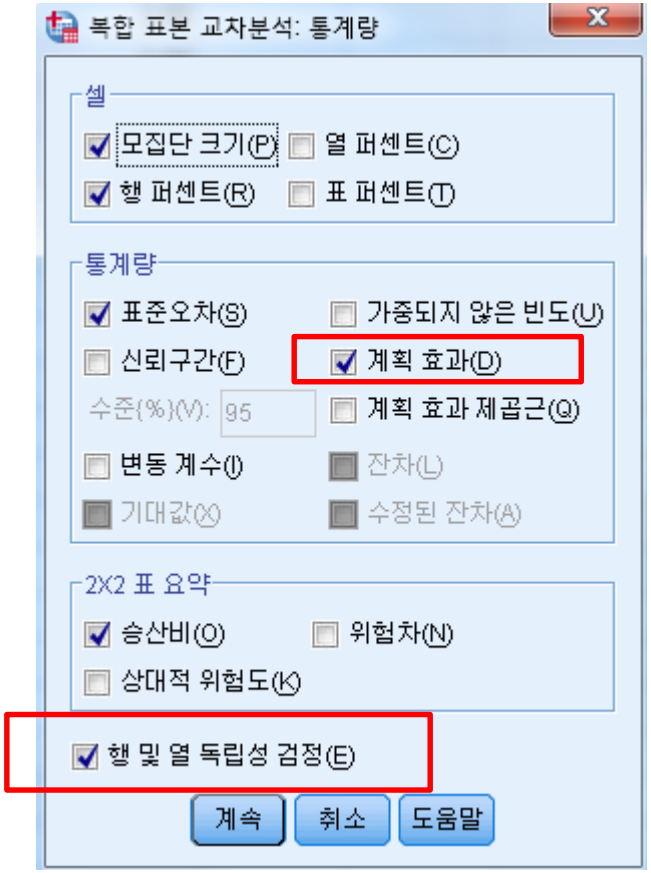
<비고> 가중치 부여 결과
- 모수 추정치 일치(비편향)
- 표준오차 추정치 편향

[2] SPSS 복합표본: 카이제곱검정

- **복합표본 카이제곱검정**
  - 1단계 : 복합표본설계 설정
  - 2단계 : 복합표본 교차분석



– 복합표본 교차분석의 통계량



각 셀에서 추정을 위한 설계효과를 계산

$$Deff = \frac{Var_{CS}}{Var_{SRS}} \qquad Deft = \frac{SE_{CS}}{SE_{SRS}}$$

Rao-Scott test statistic 계산

$$\frac{X_F^2}{(r-1)(c-1)} \approx \frac{X^2}{E\left(X^2\right)} \sim F\left((r-1)(c-1), (r-1)(c-1)K\right)$$

- 참고 : 설계 효과
  - 의미 : SRS분산과 복합표본설계의 분산을 비교하여 분산의 과소/과대 추정을 검토하는데 이용
  - 설계효과의 이용
    - SRS 분산을 이용한 복합표본의 분산 : 설계효과 x SRS 분산
    - 효과적인 표본크기의 계산 : **Effective sample size**
      - **복합표본크기의 분산을 기준으로 볼 때, 동일한 수준의 분산을 얻기 위해 필요한 SRS 설계에서의 표본크기를 의미함**

$$effective \quad size = \frac{n}{deff}$$

- **참고 : 가중치를 활용한 효과적인 표본크기의 계산**

$$effective \quad size = \frac{\left(\sum w_i\right)^2}{\sum {w_i}^2}$$

SPSS 복합분석 : 교차분석

- **층화설계 무시, 가중값 미반영**

			gr_y		전체
			1.00	2.00	
gr_x	1.00	빈도	11	1	12
		gr_x 중 %	.92	.08	1.00
	2.00	빈도	1	11	12
		gr_x 중 %	.08	.92	1.00
전체		빈도	12	12	24
		gr_x 중 %	.50	.50	1.00

- **층화설계 무시, 가중값 반영**

			gr_y		전체
			1.00	2.00	
gr_x	1.00	빈도	107	12	119
		gr_x 중 %	.90	.10	1.00
	2.00	빈도	8	114	122
		gr_x 중 %	.07	.93	1.00
전체		빈도	115	126	241
		gr_x 중 %	.48	.52	1.00

- 복합표본분석 : 층화 설계 및 가중값 반영

gr_x			gr_y		
			1.00	2.00	합계
1.00	모집단 크기	추정값	106.5	12.0	118.5
		표준오차	18.416	11.489	20.309
		계획 효과	.610	1.233	.732
	gr_x 중 %	추정값	.90	.10	1.00
		표준오차	.09	.09	.00
		계획 효과	1.140	1.140	.
2.00	모집단 크기	추정값	7.5	114.0	121.5
		표준오차	6.982	20.480	20.309
		계획 효과	.714	.746	.732
	gr_x 중 %	추정값	.06	.94	1.00
		표준오차	.06	.06	.00
		계획 효과	.722	.722	.
합계	모집단 크기	추정값	114.0	126.0	240.0
		표준오차	18.604	18.604	.000
		계획 효과	.616	.616	.
	gr_x 중 %	추정값	.48	.53	1.00
		표준오차	.08	.08	.00
		계획 효과	.616	.616	.

독립성 검정 결과

- **설계무시**
  - 설계 무시/가중값 반영 분석은 모집단크기 기분 분석으로 적절하지 않은 결과를 제공함에 주의

	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	16.667	1	.000
연속수정b	13.500	1	.000
우도비	19.503	1	.000
유효 케이스 수	24		

	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	167.786	1	.000
연속수정b	164.462	1	.000
우도비	196.730	1	.000
유효 케이스 수	241		

- **복합표본**

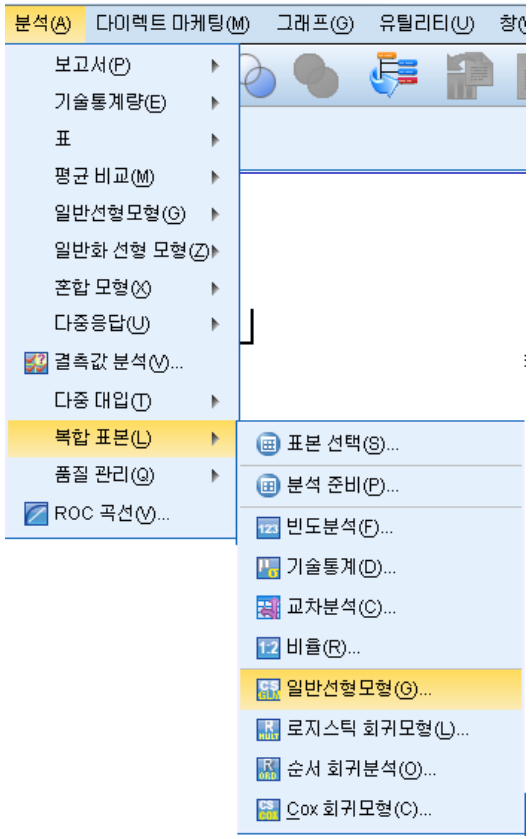
독립성 검정

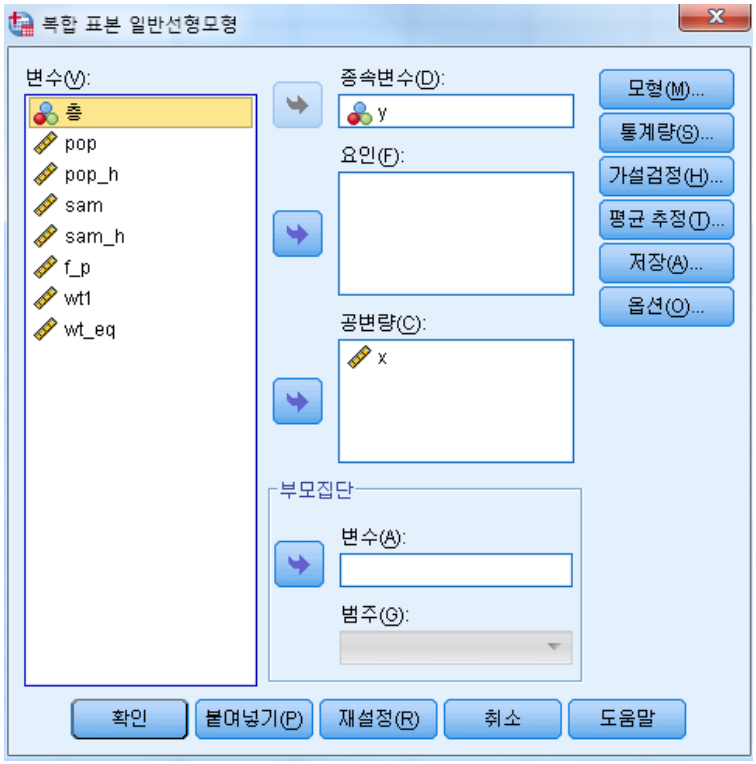
		카이 제곱	수정된 F	df1	df2	유의확률
gr_x * gr_y	Pearson	16.853	18.744	1	21	.000
	우도비	19.811	22.033	1	21	.000

  - 수정된 F는 수정된 2차 라오-스캇 카이제곱 통계량을 의미

[3] SPSS 복합표본 : 회귀분석

- 복합표본조사자료분석
  - 일반선형모형 : 분산분석, 회귀분석, 공분산분석
  - 분석을 위해 모형, 통계량, 가설검정, 평균 추정 설정





- SPSS 복합표본 회귀분석 결과

	평균
종속 변수 y	91.652
공변량 x	85.164

모형 요약a

R 제곱	.929

모형 효과 검정a

소스	df1	df2	Wald F	유의확률
(정확 모델)	1.000	21.000	187.931	.000
(절편)	1.000	21.000	17.033	.000
x	1.000	21.000	187.931	.000

모수 추정값a

모수	추정값	표준오차	95% 신뢰구간		가설검정			계획 효과
			하한	상한	t	자유도	유의확률	
(절편)	21.892	5.304	10.861	32.922	4.127	21.000	.000	1.111
x	.8191	.0598	.695	.943	13.709	21.000	.000	1.111

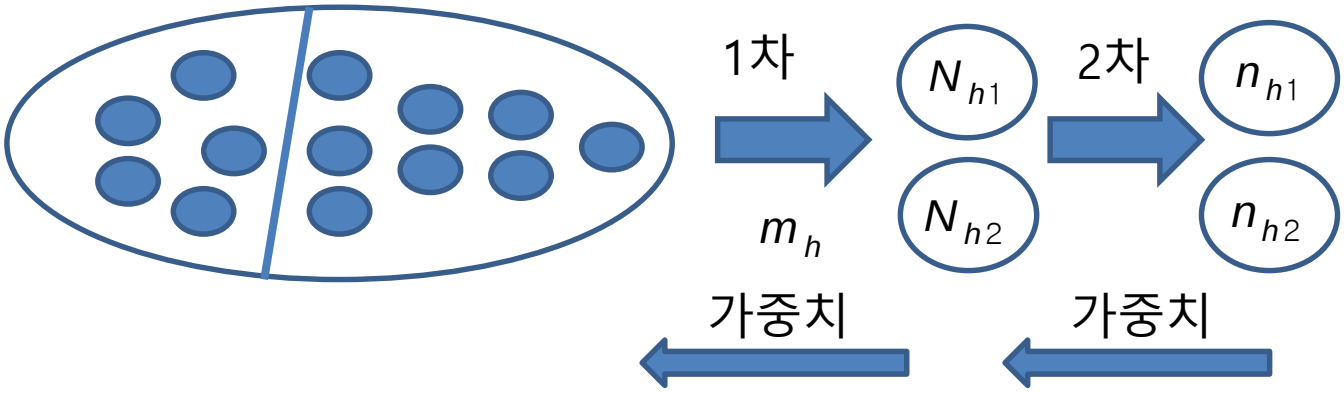
- 모수 추정 결과는 SAS와 동일
- 분산 추정값은 SAS와 SPSS의 적용 방법이 다르므로 약간 차이가 있음에 주의
- 설계효과가 1.11은 복합표본의 분산이 SRS 분산보다 1.11배 크다는 의미하므로 SRS로 분석한다면 분산이 11.1%만큼 과소 추정하게 된다는 것을 의미함

2. 전형적인 표본조사에서의 분석 이슈

- 모든 표본조사에서 복합표본설계로 수행하면 설계 특성을 모두 반영하지는 않지만 일부 주요한 표본설계 특성을 반영해 추정/분석하는 것이 일반적임
  - 가능하면 복합표본설계특성을 모두 반영하는 것이 바람직하지만 매우 복잡할 수 있어 일부 중요한 특성을 반영하는 것이 효율적(단순화의 욕구)

- 전형적인 표본 설계의 사례
  - 대부분의 실생활 조사가 전형적인 조사 사례에 해당
  - 대규모 조사는 복잡한 설계로 수행되지만 완전한 조사설계를 반영한 모수와 상당히 근사한 결과를 제공하는 범위에서 분석을 위해 단순화하려는 시도가 요구

- 전형적인 표본 설계의 과정
  - 일반적으로 조사규모가 클수록 표본설계는 복잡
  - 표본추출법 : 층화집락추출법
    - 층 : 모든 모집단 개체를 L개 층으로 그룹화
    - 집락(1차 추출단위) : h층은 M_h 개 집락으로 구성해 m_h 개를 추출
  - 최종추출단위 : 표본집락마다 N_{hi} 개중 n_{hi} 개 추출(usu)



<참고> 모집단 크기를 안다면 유한모집단수정계수를 반영해 분석

결론 : 복합표본조사자료의 특성

- 대표성 확보 및 단순확률추출에 의한 표본조사의 현실적인 어려움으로 복합표본조사로 자료 수집

- 복합표본조사자료의 특성
  - **조사단위간 상이한 추출 확률**
    - ✓ 표본추출방법에 따라 추출확률이 다름
  - **무응답 표본 및 자료 발생**
    - 단위무응답
    - 항목무응답
  - **모집단 포함율의 영향**
    - 조사모집단과 대상모집단과의 차이
    - 표본틀의 영향
    - 표본설계과정에서 고려하지 않은 중요변수의 영향

- 복합표본조사설계에서 주로 이용되는 표본추출법
  - 층화추출법
  - 집락추출법
  - 다단계 추출법
  - 혼합추출 : 전수조사층+표본조사층

- 참고 : 표본단위의 추출확률이 균등한 표본설계
  - 단순확률추출
  - 비례배분의 층화추출
  - 규모크기의 확률 비례 추출(PPS sampling)

- 복합표본조사자료를 단순확률표본으로 분석할 때 발생되는 문제
  - 단순확률표본 가정에 위배
  - 표준의 통계패키지를 이용한 분석이 어려움
  - 가중치 미반영시 편향된 결과 산출
  - 분산 과소 추정
    - 비선형 추정의 문제 발생

- 복합표본조사자료분석의 고려 사항
  - 변동 추출 확률 : 가중값으로 반영 (varying selection probability)
  - 표본설계(sample design)
    - 층(stratification)
    - 집락(cluster)
  - 분산 추정(variance estimation)
- 가중치를 부여하는 이유
  - 불균등 추출 확률을 보정하기 위해
  - 무응답을 보정하기 위해
  - 사후 층화(post-stratification)

    - 알려진 모집단 분포(성, 연령 등의 분포 혹은 규모 등의 일치)와 표본 분포를 일치시키기 위해 조정하기 위해

– 추정 및 분석을 위한 통계 패키지의 활용

- 복합 표본설계를 반영한 분석이 필요
  - 예 : 2단계 층화집락추출법
  - 1차 추출(psu) : 층마다 행정구역 pps 추출
  - 2차 추출(ssu) : 표본 행정구역내에서 조사 대상자를 확률 추출

- 통계패키지 사용에 주의
  - 일반통계패키지 : SRS 및 균등 추출확률 가정
    - 복합표본조사분석 시 결과 특성
      - 추정치 : 편향
      - 분산추정치 : 과소추정

  - 복합표본조사분석이 가능한 전문 통계패키지 사용이 요구
    - 설계특성 반영 : 층, 집락
    - 추출과정 : 1차 추출, 2차 추출
    - 가중치 : 단계별 가중치 및 설계 가중치, 최종 가중치 등
    - 분산 추정 방법 반영

– 설계효과(design effect : deff)의 평가(혹은 검토/측정)

$$deff = \frac{Var_{design}}{Var_{SRS}} = \left(\frac{se_{design}}{se_{SRS}} \right)^2$$

- 설계효과의 개념
  - 복합표본설계의 효과를 표현하는 측도
  - SRS 분산과 복합표본설계의 분산 비로 계산
  - 일반적으로 deff>1
  - 집락내 동질성으로 인한 분산 팽창(variance inflation) 측도로 이용

    $deff \approx 1 + \delta_x(\bar{n} - 1)$ where δ_x : ICC, $\bar{n}$: 집락당 평균 표본크기
    - ICC는 집락내 대상들이 매우 이질적이면 <0, 매우 동질적이면 1에 근사
    - 참고 : 회귀계수의 설계효과 $deff \approx 1 + (\bar{n} - 1)\delta_y \delta_x$
- 설계효과의 해석
  - 복합표본설계를 무시하고 SRS로 추정한다면 설계효과만큼 과소 혹은 과대 추정되는 것으로 해석
  - SRS 표본크기 기준의 표본크기 효과를 보여주는 측도로도 해석
  - 예 : deff=1.85라면, 복합표본설계를 무시하고 SRS로 분석한다면 표본분산을 85%만큼 과소추정하게 되므로 이를 이용해 신뢰구간을 계산하면 구간 폭이 좁아지게 됨으로 설명