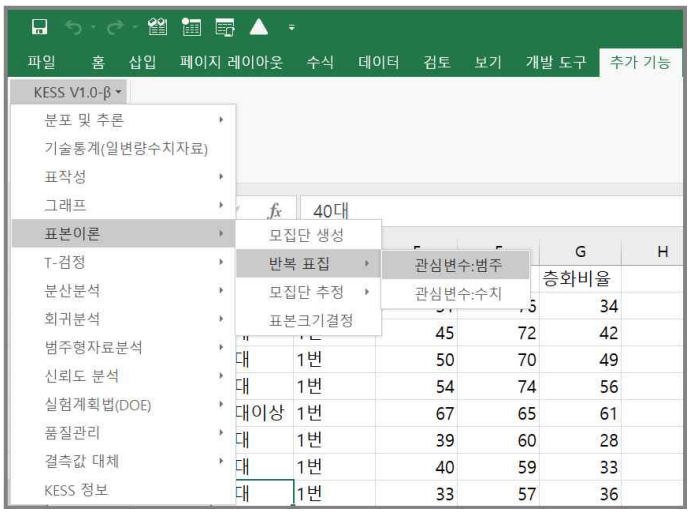
◉ 표본이론: KESS_Sampling

○ 표본이론 > 반복 표집 > 관심변수: 범주

◦ 관심변수가 범주형 자료일 때 다양한 표집방법(단순확률표집, 층화표집, 계통표집)으로 표집을 반복하여 출력해주는 동시에 관련 통계값을 정리하여 제고함



◦ 반복표집을 하기 위해서는 먼저 모집단 또는 모집단을 정리한 시트가 필요함. 이 시트 상에서 SRS, 층화확률표집, 계통표집으로 표본 추출하는 과정을 반복하여 출력해 줌(집락표집은 이후에 추가할 예정임). 기본 설정은 '관심변수'에는 표집하고자 하는 변수를 지정하며 이 변수는 반드시 있어야 함. '특성변수'에는 같이 표시하고 싶은 변수를 지정함. 모집단을 정리 시트인 경우 '비율변수'에 관심변수(특성변수 포함)의 값(그룹)이 차지하는 가중변수를 지정하고, '응답율변수'에는 해당 행에서의 자료에 대한 응답율이 있는 경우 응답율 관련 변수를 지정함.

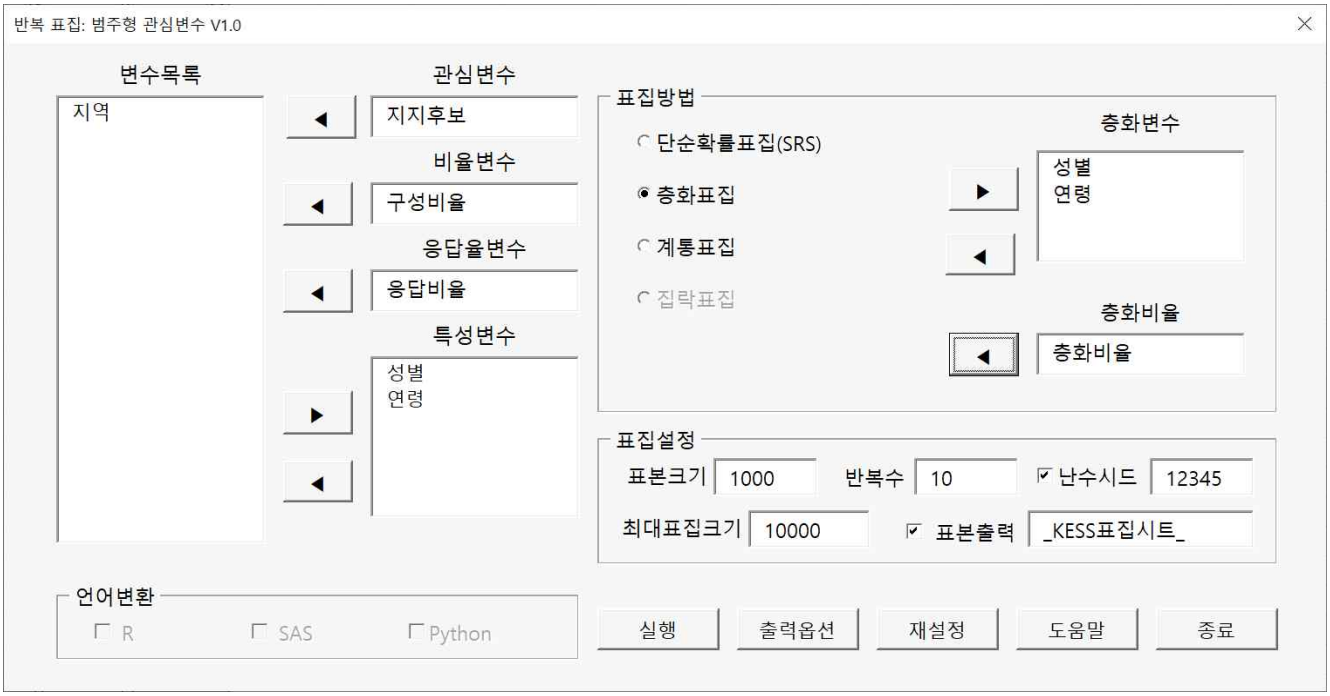
◦ '표집설정'에서 표본크기와 반복수를 설정하며 반복수는 최대 1000번까지 수행할 수 있도록 하였음. 난수시드를 사용할 수 있고 '응답율 변수'가 지정된 경우 표집과정이 늘어날 수 있어 최대표집크기를 지정함. 표집결과는 '표본출력'에 지정된 시트로 출력되며 '표본출력'이 활성화되지 않은 경우에는 표집의 결과(관심변수의 범주 비율)를 결과시트에 출력해 줌. '표본출력'에 지정된 시트이름이 기존 시트이름과 같은 경우 삭제여부를 확인하고 실행함

【단순확률표집】

◦ 단순확률표집(SRS)은 위의 공통부분 이외 특별히 설정할 내용은 없음

◦ '응답율변수'가 지정되어 있는 경우 결과출력에는 응답자만 분석한 결과와 무응답자가 포함된 전체 모집단에 대해 범주의 비율을 표시해 주고 각 반복 표집의 결과들을 동일한 내용으로 출력해줌

【층화확률표집】



○ 층화표집을 선택하면 오른쪽에 '층화변수'와 '층화비율'을 선택할 수 있는 박스가 생성되며 '층화변수'는 왼쪽의 '특성변수' 리스트에 있는 변수를 선택하고 '층화비율'은 '변수목록'에서 선택하도록 함

○ 결과시트에는 각각의 반복에서 목표한 표본의 대한 범주별 비율과, 최대표집크기 내에서 각 층에서 목표된 인원만큼 표집하는 과정에서 이미 표집이 완료되어 제외한 자료를 포함한 범주별 비율, '응답율변수'가 지정되어 있는 경우 무응답자까지 포함했을 때의 범주별 비율을 같이 출력해 줌. 모집단에 대해서는 응답자만 적용했을 때와 무응답자를 포함했을 때의 비율을 제공함

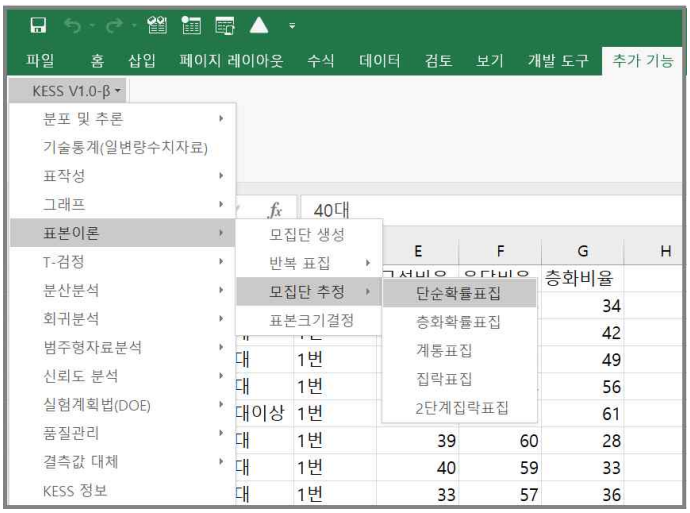
【계통표집】

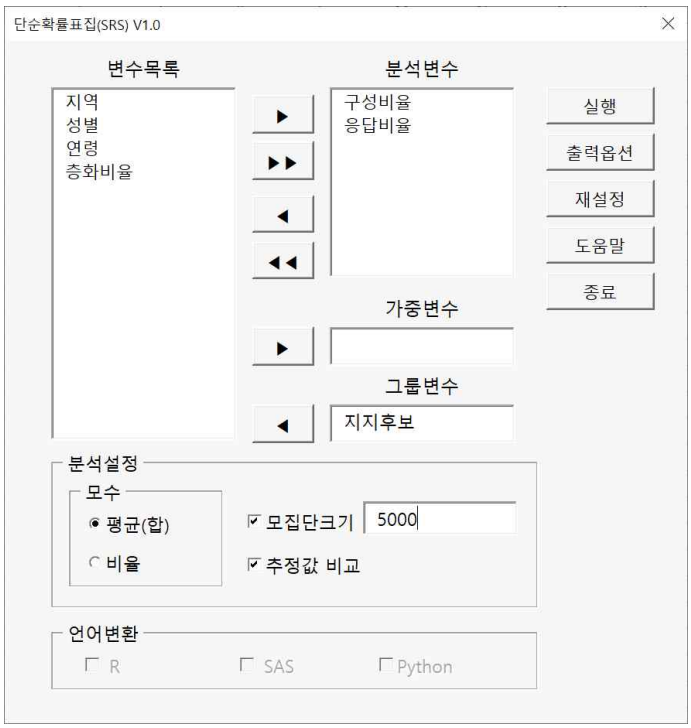
○ 계통표집을 선택하면 오른쪽에 추출간경을 설정하는 박스가 생성됨. 계통표집의 경우 모집단을 정리한 시트가 아닌 표집틀(모집단 리스트) 시트가 필요하기 때문에 비율변수는 사용할 수 없으며 표집틀은 "표본이론 > 모집단생성"을 통해 만들 수 있음.

○ 결과시트에는 "추출간격"을 바탕으로 무작위로 선택된 초기위치와 표집결과를 출력해 주고 '응답율변수'가 지정되어 있는 경우 무응답자를 포함한 결과도 함께 제공함

○ 표본이론 > 모집단 추정 > 단순확률표집

○ 단순확률표집(Simple Random Samping, SRS) 하에서의 모수(평균, 총계, 비율)에 대한 추정값, 추정값의 분산, 추정오차한계를 계산해 줌(참고문헌: 표본조사의 이해와 활용 6판, 김영원 외 3인 역).





○ 분석할 변수들을 '분석변수' 리스트에 지정함. 모수가 "평균(합)"인 경우 분석변수는 수치자료이어야 함. 자료의 빈도를 나타내는 변수가 별도로 있는 경우 '가중변수'를 지정함. 그룹별로 추정모수를 비교하고자 하는 경우 분석설정에서 "추정값 비교"를 선택하고 '그룹변수'에 해당변수를 지정함.

○ "추정값 비교"의 경우 모수가 "평균(합)"이면 각 변수에 대해 그룹별 요약 통계값과 모든 그룹별로 다중비교를 실시함. 모수가 "비율"인 경우 각 변수 내의 범주 간 비율을 비교(공분산 반영)한 결과와 각 변수에 대해 범주별로 지정한 그룹변수의 그룹 간 비율을 비교(독립)한 결과를 제공하고 만약 '그룹변수'를 지정하지 않고 "추정값 비교"를 활성화하면 각 변수 내의 범주 간 비율을 비교한 결과만 제공함. (분석예제 잠고)

○ "모집단크기"를 활성화하지 않으면 무한모집단으로 처리함.

【평균(합) 선택】

단순확률표집: Sheet1
유한모집단: N =5000

변수	표본크기	중앙값	표준편차	평균			총계		
				추정치	추정치분산	추정오차한	추정치	추정치분산	추정오차한계
구성비율	30	39.500	23.110	36.900	17.696	8.413	184500.000	442407120.690	42066.952
응답비율	30	71.500	11.186	71.967	4.146	4.072	359833.333	103654967.433	20362.217

그룹별 요약 통계값

변수	지지후보	표본크기	평균	중앙값	표준편차
구성비율	1번	10	51.200	47.500	17.962
	2번	10	47.900	45.500	17.104
	3번	10	11.600	11.500	5.739
응답비율	1번	10	63.100	62.500	10.268
	2번	10	72.500	70.500	7.735
	3번	10	80.300	81.000	8.551

추정값 비교

변수	비교그룹	평균 차	표준오차	하한	상한
구성비율	1번 - 2번	3.300	7.843	-12.387	18.987
	1번 - 3번	39.600	5.963	27.674	51.526
	2번 - 3번	36.300	5.705	24.890	47.710
응답비율	1번 - 2번	-9.400	4.065	-17.531	-1.269
	1번 - 3번	-17.200	4.226	-25.651	-8.749
	2번 - 3번	-7.800	3.646	-15.093	-0.507

【비율 선택】

단순확률표집: Sheet1
유한모집단: N =10000

변수	표본크기	모비율				
		범주	빈도	추정치	추정치분산	추정오차한계
성별	1107	남	576	0.520	0.000	0.028
		여	531	0.480	0.000	0.028
연령	1107	20대	186	0.168	0.000	0.021
		30대	226	0.204	0.000	0.023
		40대	235	0.212	0.000	0.023
		50대	244	0.220	0.000	0.024
		60대이상	216	0.195	0.000	0.022

변수 내 범주 간 비율비교

변수	표본크기	비교범주	비율차	표준오차	하한	상한
성별	1107	남 - 여	0.041	0.030	-0.019	0.101
연령	1107	20대 - 30	-0.036	0.018	-0.073	0.000
		20대 - 40	-0.044	0.018	-0.081	-0.007
		20대 - 50	-0.052	0.019	-0.090	-0.015
		20대 - 60	-0.027	0.018	-0.063	0.009
		30대 - 40	-0.008	0.019	-0.047	0.031
		30대 - 50	-0.016	0.020	-0.055	0.023
		30대 - 60	0.009	0.019	-0.029	0.047
		40대 - 50	-0.008	0.020	-0.048	0.031
		40대 - 60	0.017	0.019	-0.021	0.056
		50대 - 60	0.025	0.019	-0.013	0.064

범주별 그룹의.빈도

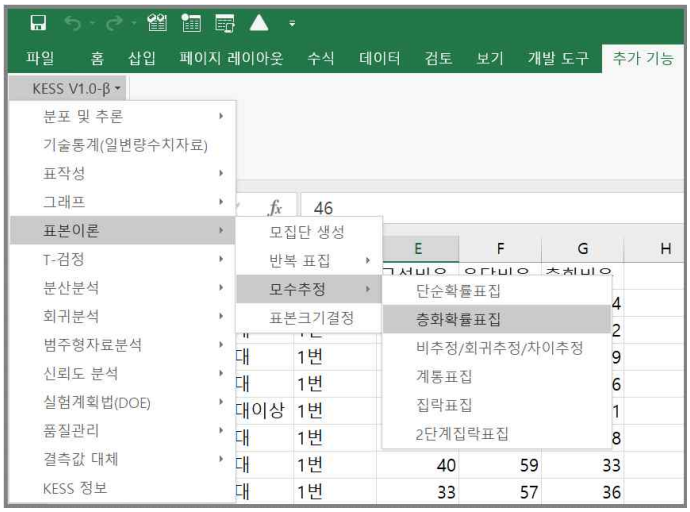
변수	범주	전체	1번	2번	3번
성별	남	576	247	276	53
	여	531	265	203	63
	소계	1107	512	479	116
연령	20대	186	70	87	29
	30대	226	85	105	36
	40대	235	83	125	27
	50대	244	118	111	15
	60대이상	216	156	51	9
	소계	1107	512	479	116

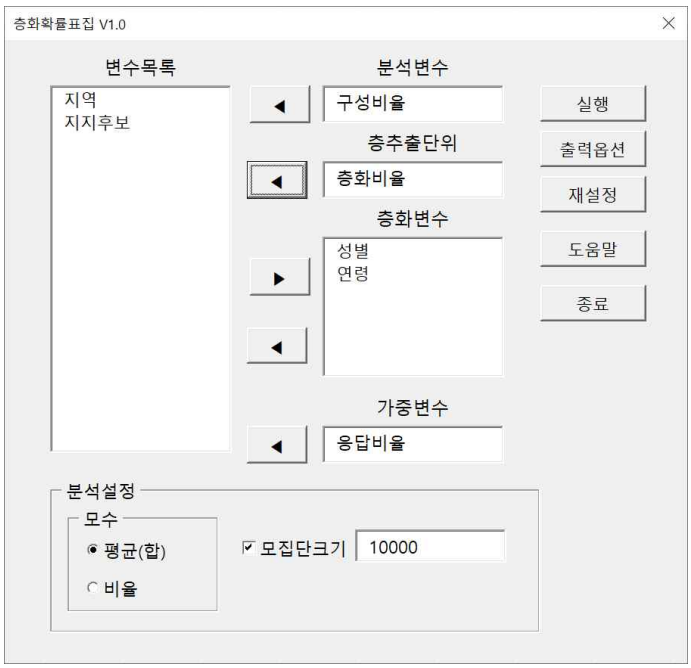
범주별 그룹 간 비율비교

변수	범주	1번 - 2번				1번 - 3번				2번 - 3번			
		비율차	표준오차	하한	상한	비율차	표준오차	하한	상한	비율차	표준오차	하한	상한
성별	남	-0.094	0.032	-0.157	-0.031	0.026	0.051	-0.077	0.128	0.119	0.051	0.016	0.222
	여	0.094	0.032	0.031	0.157	-0.026	0.051	-0.128	0.077	-0.119	0.051	-0.222	-0.016
연령	20대	-0.045	0.023	-0.091	0.002	-0.113	0.043	-0.199	-0.027	-0.068	0.044	-0.156	0.019
	30대	-0.053	0.025	-0.103	-0.003	-0.144	0.046	-0.236	-0.052	-0.091	0.047	-0.185	0.003
	40대	-0.099	0.026	-0.151	-0.047	-0.071	0.042	-0.156	0.014	0.028	0.044	-0.060	0.116
	50대	-0.001	0.027	-0.055	0.052	0.101	0.036	0.029	0.174	0.102	0.037	0.029	0.176
	60대이상	0.198	0.025	0.149	0.248	0.227	0.032	0.163	0.291	0.029	0.029	-0.028	0.086

○ 표본이론 > 모집단 추정 > 층화확률표집

◦ 층화확률표집(Stratified Samping) 하에서의 모수(평균, 총계, 비율)에 대한 추정값, 추정값의 분산, 추정오차한계를 계산해 줌(참고문헌: 표본조사의 이해와 활용 6판, 김영원 외 3인 역).





◦ 분석할 변수를 '분석변수', 층의 모집단 비율 또는 빈도를 층추출단위에 지정함. "각 자료에 가중치가 있는 경우 '가중변수'에 지정하고 층화변수는 '층화변수' 리스트에 지정함

◦ 체크박스 "모집단크기"가 선택된 경우 '층추출단위'의 변수는 비율로 반영하여 모집단크기 N에 맞게 빈도를 배분하며 선택되지 않은 경우 층추출단위의 합을 모집단크기 N으로 사용함

◦ 분석설정의 모수에서 "평균(합)"을 선택하면 아래 분석예제에서 보는 것처럼 각 층화별로 층화크기(N_i), 층화비율(N_i/N), 표본크기(n_i)와 함께 평균, 표준편차, 평균의 표준오차, 평균의 신뢰구간, 총계, 총계의 표준오차를 제공함. 아래 쪽에 전체모집단에 대한 추정결과를 제공하며 표준편차는 각 층화별 평균이 다르기 때문에 표시하지 않음

◦ 비율의 선택한 경우 '분석변수'는 범주형으로 처리하고 분석변수의 모든 범주별로 층별 추정비율, 표준오차, 신뢰구간을 제공하고 아래에 전체비율에 추정결과를 제공함

【비율 선택】

층화확률표집: 예제 5.12

범주: O

층	층화크기	층화비율	표본크기	빈도	추정비율	표준오차	신뢰하한	신뢰상한
1	155	0.500	20	16	0.800	0.086	0.629	0.971
2	62	0.200	8	2	0.250	0.153	-0.055	0.555
3	93	0.300	12	6	0.500	0.141	0.219	0.781
전체	310	1.000	40	24	0.600	0.067	0.465	0.735

범주: X

층	층화크기	층화비율	표본크기	빈도	추정비율	표준오차	신뢰하한	신뢰상한
1	155	0.500	20	4	0.200	0.086	0.029	0.371
2	62	0.200	8	6	0.750	0.153	0.445	1.055
3	93	0.300	12	6	0.500	0.141	0.219	0.781
전체	310	1.000	40	16	0.400	0.067	0.265	0.535

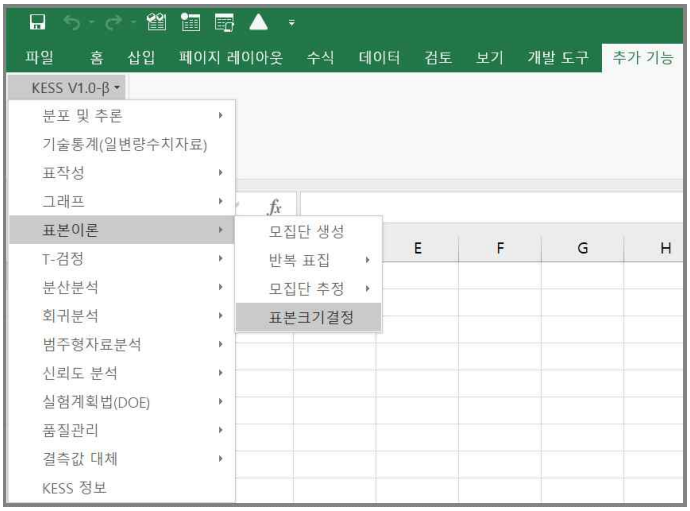
【평균(합) 선택】

층화확률표집: Sheet1

성별	연령	층화크기	층화비율	표본크기	추정평균	표준편차	표준오차(평	평균하한	평균상한	추정총계	표준오차(총계)
남	20대	801	0.080	239	29.423	14.108	0.764	27.894	30.951	23567.498	612.267
	30대	989	0.099	241	39.299	18.834	1.055	37.189	41.409	38866.469	1043.464
	40대	1152	0.115	237	47.709	27.340	1.583	44.543	50.874	54960.608	1823.306
	50대	1317	0.132	240	39.342	25.848	1.509	36.324	42.359	51812.975	1987.083
	60대이상	1436	0.144	219	28.123	26.929	1.675	24.773	31.474	40385.041	2405.573
여	20대	658	0.066	218	31.752	11.175	0.619	30.514	32.990	20893.009	407.259
	30대	777	0.078	204	35.074	11.018	0.662	33.749	36.398	27252.132	514.730
	40대	848	0.085	199	28.744	14.227	0.882	26.979	30.508	24374.673	748.175
	50대	964	0.096	192	36.391	23.795	1.537	33.317	39.464	35080.563	1481.455
	60대이상	1058	0.106	170	34.682	32.858	2.309	30.065	39.300	36693.929	2442.665
전체		10000	1.000	2159	35.391		0.487	34.418	36.365	353912.707	4866.994

○ 표본이론 > 표본크기결정

○ 유한모집단에서 표집방법과 모수에 따른 표본크기를 계산해 줌(참고문헌: 표본조사의 이해와 활용 6판, 김영원 외 3인 역).



○ 표집방법 콤보박스에서 표집방법을 선택함. '단순확률표집', '층화확률표집', '계통표집', '집락표집', '비추정' 중 하나를 선택

○ 관심모수는 "평균", "총계", "비율" 중 하나를 선택할 수 있으며 표집방법과 관심모수에 따라 필요정보의 레이블이 변경됨

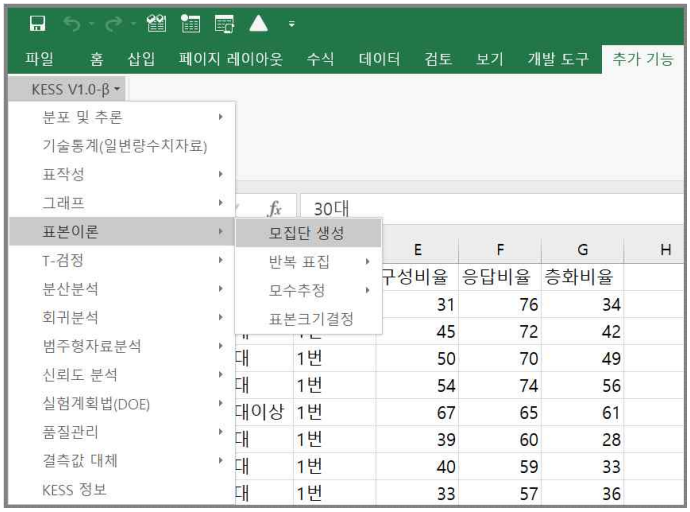
○ 표집방법에서 층화확률표집을 선택한 경우, 층설정 프레임이 활성화되고 표본배분을 위해 층비율, 관측비용, 비례배분 중 하나를 선택해야 함. 층비율, 관측비용, 층추출단위수, 산포(분산,표준편차,...)는 층의 수만큼 "/" 또는 ","로 분리하여 입력해야 하며 값이 한 개만 입력되는 경우 동일한 값이 적용됨

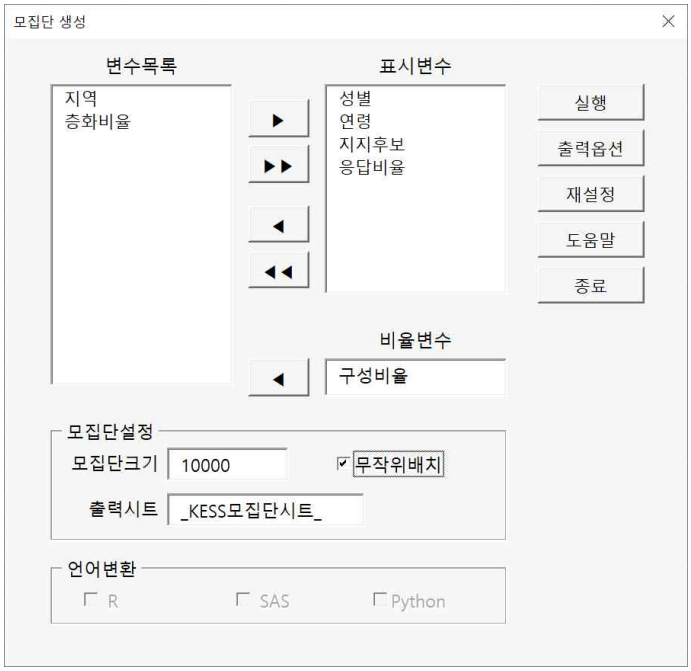
○ "비추정"의 경우 관심모수의 "비율"은 참고문헌에서 'R'로 표시한 것임

○ 결과 프레임의 출력값의 D, 분자, 분모는 참고문헌에서 표본크기를 계산하는데 사용된 수식의 값을 의미하며 층화확률표집에서의 표본배분 결과는 층설정에서 선택된 층비율, 관측비용, 비례배분에 따른 값임. 참고문헌에서의 일부 결과가 다른 것은 참고문헌에서의 소수점 처리가 세밀하지 못한 것으로 이 프로그램의 결과가 더 정확한 것임

○ 표본이론 > 모집단 생성

◦ 구성 비율 정보에 맞게 유한모집단을 만들어 줌





◦ 시트에 표시하고 싶은 변수를 '표시변수'에 지정하고 구성비율 정보를 '비율변수'에 지정함

◦ 모집단 설정에서 모집단 크기과 출력할 시트를 설정하고 자료를 무작위로 배치하고 싶으면 "무작위배치" 체크 박스를 선택함

	A	B	C	D	E	F
1	번호	성별	연령	지지후보	응답비율	
2	1	남	60대이상	2번	68	
3	2	남	30대	2번	84	
4	3	남	40대	2번	79	
5	4	여	40대	1번	57	
6	5	여	40대	1번	57	
7	6	여	30대	2번	70	
8	7	남	20대	1번	76	
9	8	여	30대	2번	70	
10	9	여	40대	2번	69	
11	10	여	20대	2번	79	
12	11	남	30대	1번	72	
13	12	여	60대이상	1번	43	
14	13	여	50대	2번	66	
15	14	여	40대	2번	69	
16	15	여	50대	1번	55	

... _KESS모집단시트_ _KESS표집시트_ 예제 4.7 예제 4.4