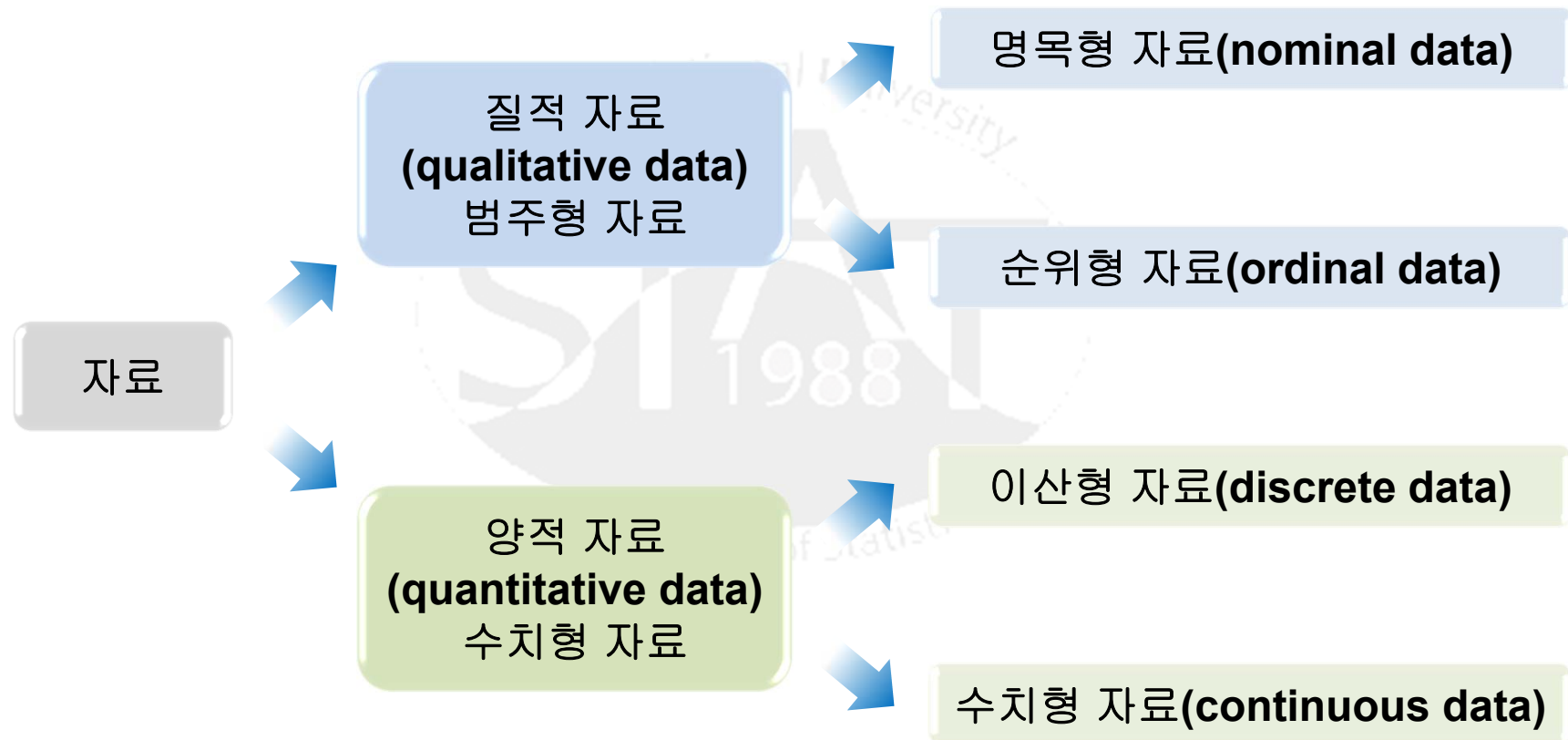


PASW Workshop
PART 4-1

범주형 자료분석

❖ 범주형 자료의 정의



범주형 자료분석

❖ 자료 유형에 따른 통계 분석

종속변수 (반응변수)	독립변수 (설명변수)	통계분석법
수치형	범주형(2개 범주)	T-검정, paired T-검정
수치형	범주형(3개 이상)	분산분석(ANOVA)
범주형	범주형	χ^2 -검정, 로짓분석
수치형	수치형	회귀분석
수치형	혼합(수치형+범주형)	공분산분석(ANCOVA)
범주형	혼합(수치형+범주형)	로짓분석
생존시간	혼합(수치형+범주형)	생존분석

범주형 자료분석

❖ 예제

- 돌발성 난청의 치료예후에 영향을 미칠 것으로 의심되는 인자들의 유의성을 평가하기 위하여 돌발성 난청으로 치료를 받은 253명의 환자들을 대상으로 난청 발생 시 나이 및 임상 특성, 치료 후 회복 여부 등을 조사하였다. 이명 동반 여부와 어지럼증 동반 여부간에는 관련성이 있는가?

		어지럼증		합계
		없음	동반	
이명	없음	27	9	36
	동반	121	96	217
합계		148	105	253

$$\Rightarrow p_{11} = 27 / 36 = 75.0\%$$

$$\Rightarrow p_{12} = 121 / 217 = 55.8\%$$

- 결과적으로 비율의 차이가 생겼음

범주형 자료분석

❖ 예제

		어지럼증		합계
		없음	동반	
이명	없음	27	9	36
	동반	121	96	217
합계		148	105	253

관측도수 : O_{11}

		어지럼증		합계
		없음	동반	
이명	없음	E_{11}	E_{12}	36
	동반	E_{21}	E_{22}	217
합계		148	105	253

기대도수 : $E_{ij} = np_{ij} = \frac{n_{i0}n_{0j}}{n}$

$$E_{11} = 36 \times \frac{148}{253} = 21.1$$

$E_{11} = 21.1$	$E_{12} = 14.9$
$E_{21} = 126.9$	$E_{22} = 90.1$

범주형 자료분석

❖ 비율에 대한 추론 – 카이제곱 (chi-square test) 검정

- 귀무가설 (H_0) : 이명과 어지럼증은 서로 독립이다
= 이명과 어지럼증 간에 관련성이 없다
- 귀무가설이 참이라면
- 각 셀의 관측도수(O_{ij})와 기대도수(E_{ij}) 차이는 근소
- 검정통계량

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- χ^2 값이 크면 귀무가설 기각
 - 통계분포 이론 : 귀무가설이 참이면 은 자유도 d 인 카이제곱 분포를 따름
(n 이 충분히 클 때)
 - 자유도의 계산 : (행 범주 수 - 1)X(열 범주 수 - 1)

범주형 자료분석

❖ 카이제곱 검정 (chi-square test) 유형

- 적합도 검정 (일표본)
 - 각 범주의 비율이 실제 자료에 적합한지를 알아보는 검정 (goodness-of-fit test)
 - 주어진 자료가 각 범주에 속할 확률들이 제시한 값과 같은지를 검정
- 동질성 검정 (이표본)
 - 모집단들이 각 반응범주에 어떤 반응을 보이는지를 검정
- 독립성 검정 (이표본)
 - 두 가지의 특성에 따라 어떻게 분류되는지를 알아보는 검정

범주형 자료분석

❖ 적합도 검정 예

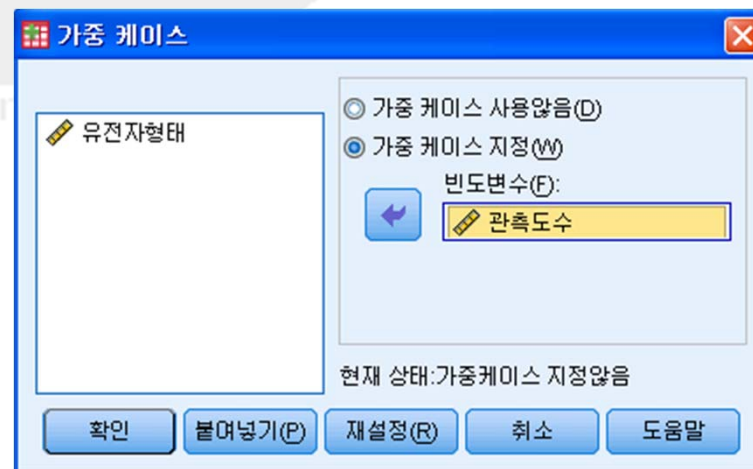
- 어떤 나무의 자가수정의 결과로 나올 수 있는 유전자의 형태는 세 종류로 분류된다고 한다. 이 세 종류를 각각 A, B, C라 할 때 생물학에서의 한 이론에 의하면 비율이 1 : 2 : 1로 나타난다고 한다. 이를 입증하기 위하여 자가수정의 결과로 생겨난 나무 100그루를 유전자의 형태로 분류하여 다음과 같은 표를 얻었다

유전자 형태	A	B	C	합계
관측도수	18	55	27	100

- 얻어진 자료로부터 생물학 이론이 제시하는 가설이 맞는지를 판단해 보자 (유의수준 5%)

❖ 분석방법

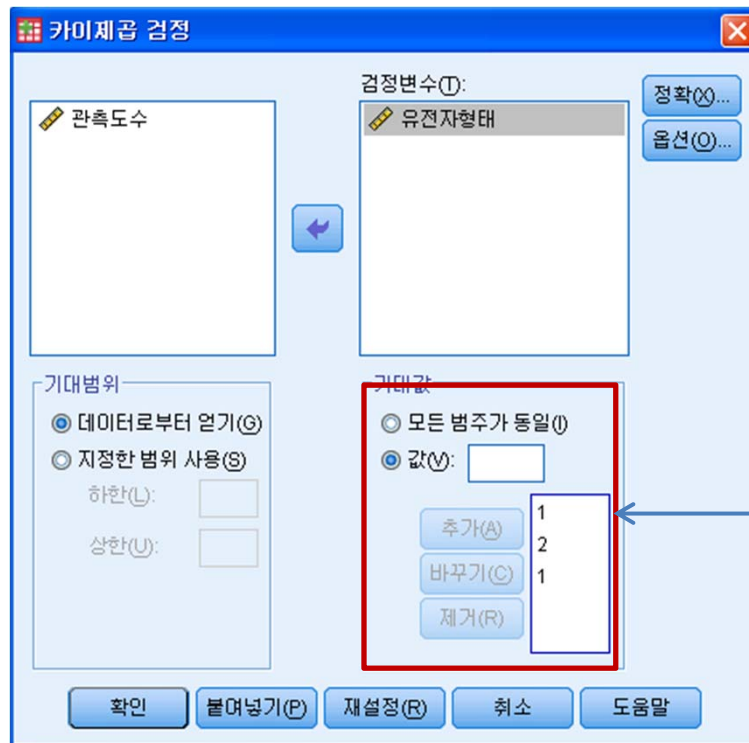
- 자료입력
- [데이터] - [가중 케이스]



범주형 자료분석

❖ 분석방법

- [분석] - [비모수적 검정] - [레거시 대화 상자] - [카이제곱검정]



비교하고자 하는 자료들의 비율을 적어준다

- 전체 케이스 수가 작은 경우에는 [정확] 단추를 눌러서 Monte Carlo 검정이나 정확한 검정을 선택할 수 있다.

범주형 자료분석

❖ 결과해석

- 첫 번째 표는 유전자 형태의 빈도표에 기대비율에 의한 기대빈도
- 관측빈도와 기대빈도의 차이인 잔차
- 두 번째 표는 카이제곱 검정 결과이다
- 검정통계량이 2.620이고 이 값에 대한 유의확률이 0.925로 나타났다.
- 따라서 생물학 생물학 이론이 제시하는 가설을 기각하지 못한다.

빈도 분석

유전자형태

	관측수	기대빈도	잔차
1,00	18	25,0	-7,0
2,00	55	50,0	5,0
3,00	27	25,0	2,0
합계	100		

검정 통계량

	유전자형태
카이제곱 ^a	2,620
자유도	2
근사 유의확률	.270

a. 0 셀 (.0%)은(는) 5보다 작은 기대빈도를 가집니다. 최소 셀 기대빈도는 25,0입니다.

범주형 자료분석

❖ 동질성 검정 예

- 두 가지 식이요법 A와 B의 효과를 비교하기 위해서 150명의 환자를 대상으로 조사를 실시하였다. 임의로 추출된 80명에게는 식이요법 A를 적용하고, 나머지 70명에게는 식이요법 B를 적용한 후 얼마간의 시간이 흐른 후에 각 환자의 건강상태에 따라 다음과 같이 세 가지 범주로 분류하였다

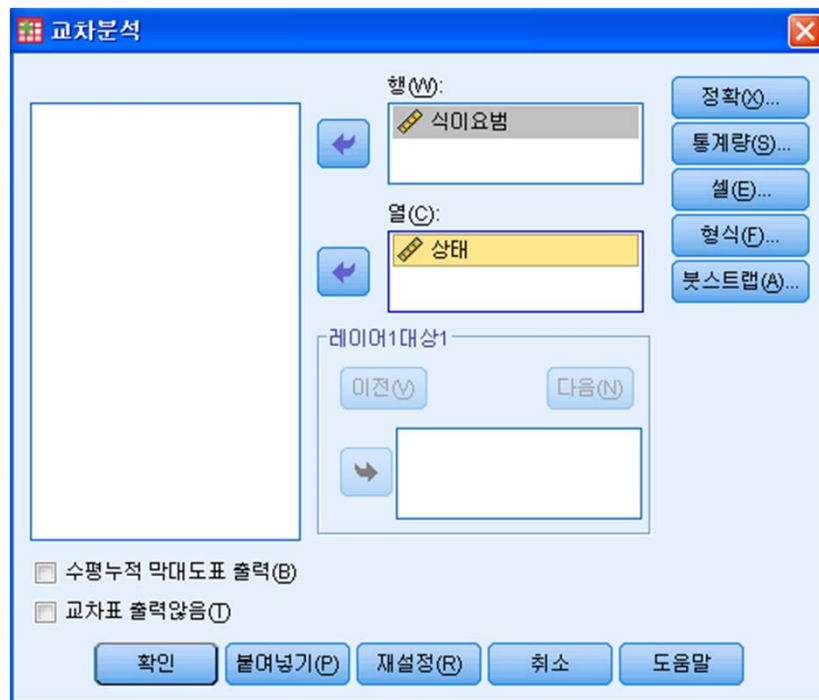
	양호	보통	불량	합계
식이요법 A	37	24	19	80
식이요법 B	17	33	20	70
합 계	54	57	39	150

- 식이요법 A와 B의 효과가 동질한지를 검정해보자 (유의수준 5%)

범주형 자료분석

❖ 분석방법

- [분석] - [기술통계량] - [교차분석]

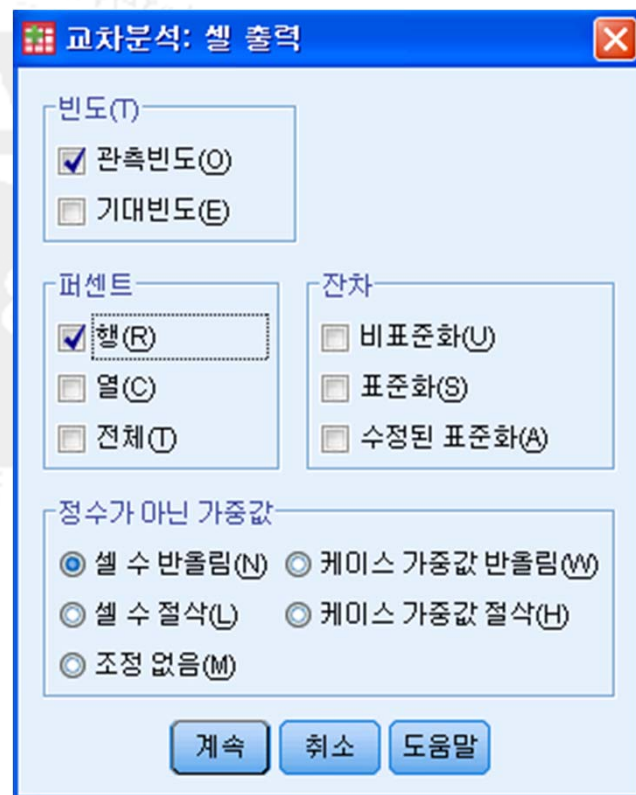
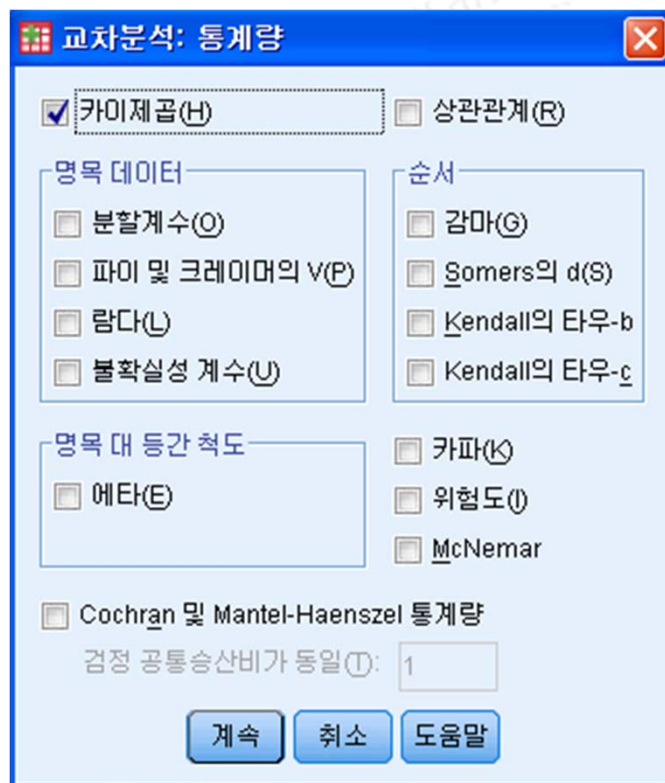


- 행과 열이 바뀌어도 상관없다

범주형 자료분석

❖ 분석방법

- 통계량을 눌러 카이제곱을 체크한다
- 셀을 눌러 퍼센트를 체크한다



범주형 자료분석

❖ 결과해석

- 첫 번째 표는 두 변수들의 교차표를 나타낸다
- 두 번째 표는 카이제곱 검정결과이다
- 검정통계량이 8.224이고 이 값에 해당하는 유의확률이 0.016이다.
- 따라서 두 식이요법은 차이가 있다

식이요법 * 상태 교차표

			상태			전체
			1,00	2,00	3,00	
식이요법	A	빈도	37	24	19	80
		식이요법의 %	46,3%	30,0%	23,8%	100,0%
B	빈도		17	33	20	70
		식이요법의 %	24,3%	47,1%	28,6%	100,0%
전체		빈도	54	57	39	150
		식이요법의 %	36,0%	38,0%	26,0%	100,0%

카이제곱 검정

	값	자유도	점근 유의확률 (양측검정)
Pearson 카이제곱	8,224 ^a	2	,016
우도비	8,372	2	,015
선형 대 선형결합	4,362	1	,037
유효 케이스 수	150		

a. 0 셀 (.0%)은(는) 5보다 작은 기대 빈도를 가지는 셀입니다. 최소 기대빈도는 18,20입니다.

범주형 자료분석

❖ 독립성 검정 예

- 마케팅담당자의 관심문제 “가족규모에 따라 구매하는 세탁기의 크기가 다른가 ?”
→ “가족규모와 구매하는 세탁기의 크기는 서로 독립적인가 아닌가 ?” ; 가족규모와 세탁기의 크기가 독립적이지 않을 때 가족규모에 따라 세탁기의 크기는 다르게 나타난다. 그러므로 “가족규모에 따라 세탁기의 크기가 다를 것이다.” 또는 “식구가 많으면 보다 큰 세탁기를 살 것이다.”가 연구가설(대립가설)이 되고, “가족규모와 세탁기의 크기는 무관하다.”라는 진술이 귀무가설로 설정된다.

	1~2명	3~4명	5명 이상	합계
소 형	25	37	8	70
중 형	10	62	53	125
대 형	5	41	59	105
합 계	40	140	120	300

- 가족규모와 세탁기 크기는 독립적이지 않는 지를 검정해보자 (유의수준 5%)

범주형 자료분석

❖ 분석방법

- 독립성 검정과 동질성 검정은 분석하는 방법이 같으므로 개인별 숙제로 남긴다



범주형 자료분석

❖ 로지스틱 회귀분석

- 범주형 자료의 종속변수에 대하여, 한 개 이상의 독립변수들을 이용하여 추측
- 일반 회귀분석과의 차이점
 - 일반회귀 분석 : 종속변수 (수치형자료)
 - 로지스틱 회귀분석 : 종속변수 (범주형 자료)
 - 이분형 로지스틱 : 종속변수 값이 두 가지인 분석
 - 다분형 로지스틱 : 종속변수 값이 세 가지 이상인 분석
- 로지스틱 회귀모형
$$\ln \left(\frac{p_y}{1 - p_y} \right) = X \beta, \quad p_y; P(y = 1),$$
$$\ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$
 - 독립변수들의 오즈비(odds ration)를 쉽게 구할 수 있는 장점이 있다
- 로지스틱 회귀분석에서의 가정
 - 분포가정 필요없다
 - 독립변수들이 다변량 정규분포를 따르면 추정결과가 보다 안정적

범주형 자료분석

❖ 로지스틱 회귀분석 예

- 연간소득과 가계지출은 주택소유에 어떤 영향을 주는가?

주택소유	연간소득	가계지출
1	4000	2000
1	3500	1800
1	5000	3500
...	...	...
...	...	...
0	3300	2200
0	3000	1700
0	2800	2000

단위 : 만원

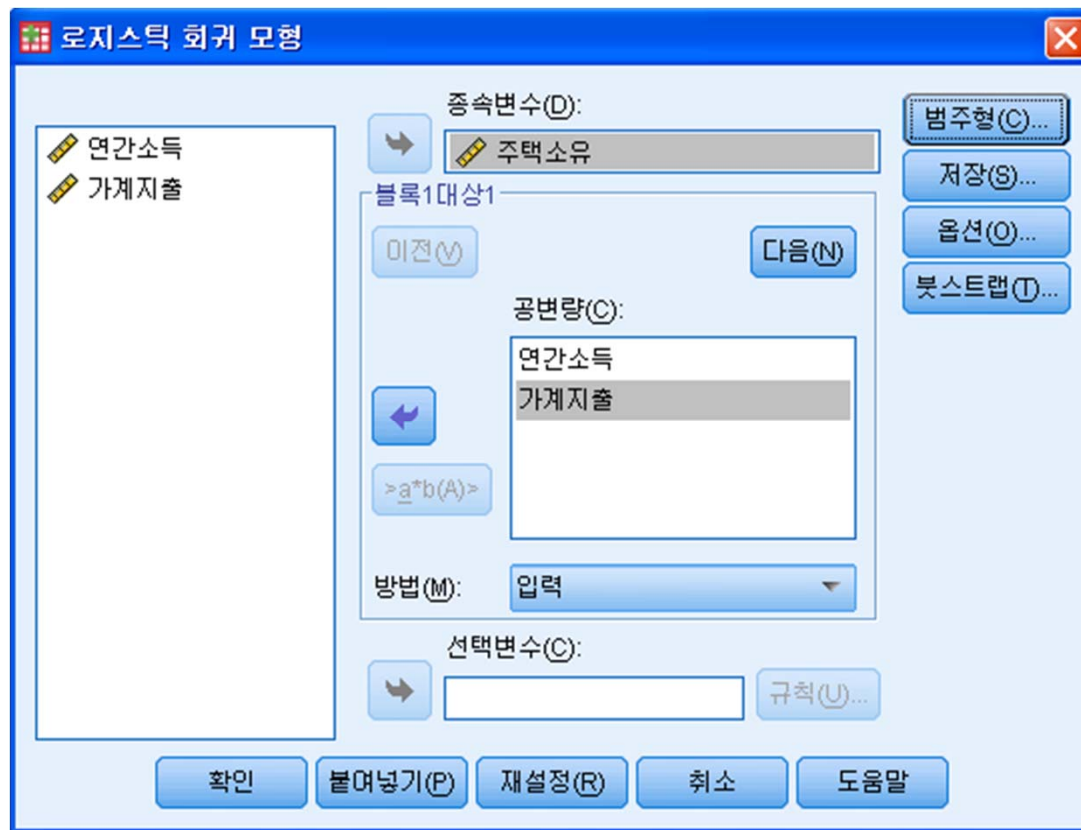
❖ 분석방법

- [분석] - [회귀분석] - [이분형 로지스틱]

범주형 자료분석

❖ 분석방법

- Y : 주택소유 ,
- X1 : 연간소득 , X2 :가계지출



- 공변량에 범주형 자료가 들어갔을 때 범주형이라는 버튼 클릭

범주형 자료분석

❖ 분석방법

- 분류도표 체크 , 신뢰구간 체크

로지스틱 회귀분석: 옵션

통계량 및 도표

☒ 분류도표(C) ☐ 추정값들의 상관계수(R)

☐ Hosmer-Lemeshow 적합도(H) ☐ 반복계산정보(I)

☐ 케이스별 잔차목록(W): ☒ exp(B)에 대한 신뢰구간(O): 95 %의 케이스 추출

☒ 밖에 나타나는 이상값(O) 2 표준편차

☒ 전체 케이스(A)

표시

☒ 각 단계마다(E) ☐ 마지막 단계에서(L)

단계선택에 대한 확률

진입(E): 0.05 제거(A): 0.10

분류 분리점(U): 0.5

최대반복계산수(M): 20

☒ 모형에 상수 포함(I)

계속 취소 도움말

범주형 자료분석

❖ 결과해석

- 변수선택 방법을 입력을 했기 때문에 한꺼번에 두 변수에 대한 모형을 분석하면 모형에 대한 유의확률이 0.004로 모형이 유의함을 알 수 있다

모형 계수 전체 테스트

		카이제곱	자유도	유의확률
1 단계	단계	19.115	2	.000
	블록	19.115	2	.000
	모형	19.115	2	.000

모형 요약

단계	-2 Log 우도	Cox와 Snell의 R-제곱	Nagelkerke R-제곱
1	36.337 ^a	.380	.507

a. 모수 추정값이 .001보다 작게 변경되어 계산반복수 6에서 추정을 종료하였습니다.

범주형 자료분석

❖ 결과해석

- 분류의 정확도를 나타내는 표입니다

분류표^a

감시될			예측		
			주탁소유		분류정확 %
			.00	1.00	
1 단계	주탁소유	.00	16	4	80.0
		1.00	6	14	70.0
전체 퍼센트					75.0

a. 절단값은 .500입니다.

- 아래 방정식에 포함된 변수 표를 통해 모형에 대한 로지스틱 회귀적합식이 완성된다

방정식에 포함된 변수

	B	S.E.	Wals	자유도	유의확률	Exp(B)
1 단계 ^a 연간소득	.002	.001	8.429	1	.004	1.002
가계지출	-.002	.001	2.435	1	.119	.998
상수항	-4.598	2.575	3.190	1	.074	.010

a. 변수가 1: 단계에 진입했습니다 연간소득, 가계지출. 연간소득, 가계지출.

$$\ln\left(\frac{p_y}{1-p_y}\right) = -4.598 + 0.002(\text{연간소득}) - 0.002(\text{가계지출})$$