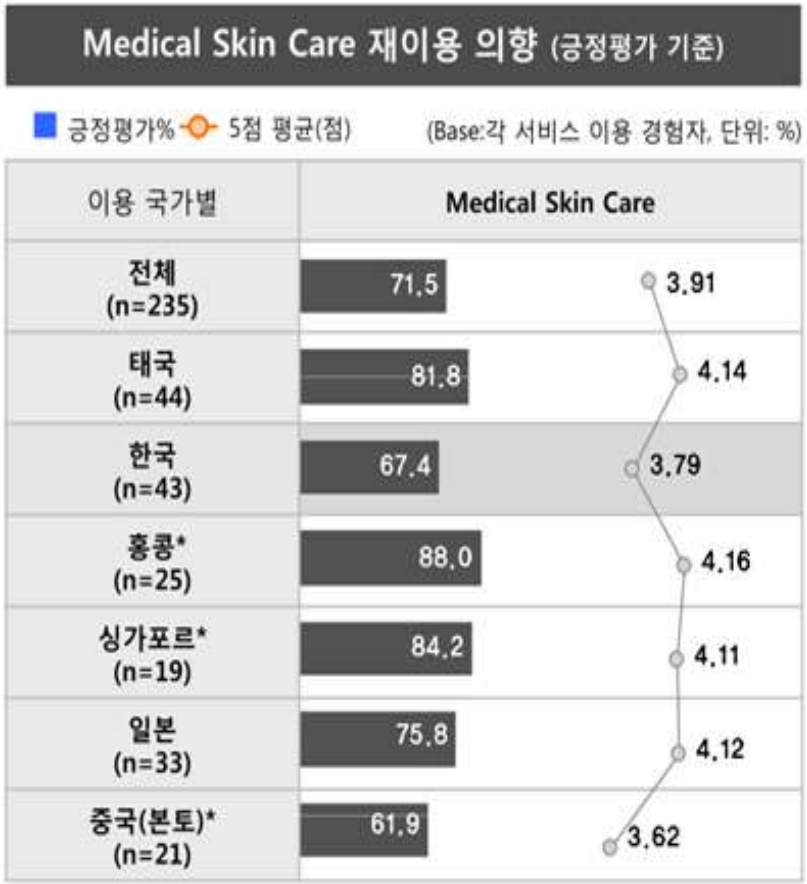
제12장 판별분석

사례: 한국 의료관광 서비스 만족도 조사



- 만족도 평가 7가지 항목 (출입국 절차 만족도, 대중교통 만족도, 숙박, 음식, 쇼핑, 관광지 매력도, 관광정보 입수 용이성)으로 재이용 의향을 알 수 있는가?

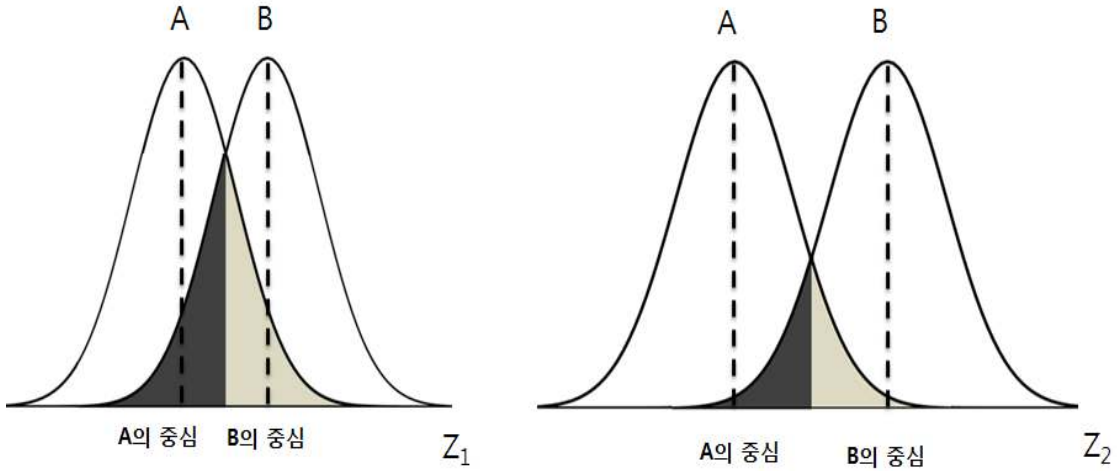
판별분석의 개요(1/3)

- 그룹 미팅 (Group blind meeting)
  - 이성 판단의 기준: 나이, 키, 얼굴, 몸매, 옷차림 (IVs)
  - 판단의 결과: 호감도 (DV)
  - 호감점수 = $a + b \times$ 나이 $+ c \times$ 키 $+ d \times$ 얼굴 $+ e \times$ 몸매 $+ f \times$ 옷차림
  - 호감여부의 판단
    - If 호감점수 ≥ 최저점수, then 호감 o
    - If 호감점수 < 최저점수, then 호감 x

- 판별분석
  - 하나의 범주적인 종속변수를 설명할 수 있는 다수의 계량적인 독립변수의 선형적인 결합을 도출해 내는 분석방법
  - 결합의 결과를 나타내는 방법 : 판별함수로 표현
  - 판별함수 (Discriminant Function)
    - $Z = W_0 + W_1X_1 + W_2X_2 + \cdots + W_nX_n$
  - 판별점수 (Discriminant Score)
    - W_i 와 X_i에 값을 대입하여 얻는 Z값

판별분석의 개요(2/3)

- 판별함수 결정의 원칙
  - 그룹간 판별점수의 중복을 최소화



(a) 판별점수의 중복이 큰 경우

(b) 판별점수의 중복이 작은 경우

  - 더 적합한 판별함수는? 중복이 작은 Z_2

판별분석의 개요(3/3)

- 판별분석의 목적
  - 데이터의 판별을 통한 예측
  - 판별에 도움이 되는 변수의 확인
  - 판별함수의 개발
  - 판별에 사용되는 독립변수의 선형관계 규명
- 판별분석의 가정
  - 각 독립변수들이 정규분포를 이루는 정규분포성
  - 각 그룹별로 동일한 분산과 공분산
  - 각 그룹별로 사례의 수가 충분히 많아야 하며 그룹간 사례 수도 비슷한 것이 좋음
- 다중회귀분석 vs. 군집분석 vs. 판별분석
  - 다수의 계량적인 독립변수들의 선형결합으로 하나의 종속변수를 예측하는 식을 찾는다는 점에서는 다중회귀분석과 유사
  - 회귀분석에서는 종속변수가 계량적인 값을 가지지만 판별분석에서의 종속변수는 범주적인 값을 가지는 차이가 있음
  - 군집분석과 판별분석 모두 그룹으로 나눈다는 측면에서는 유사성
  - 군집분석의 경우 선형적인 판별함수를 이용하기 보다 하나의 사례와 다른 사례사이의 유사성 (또는 거리)를 이용하여 그룹을 형성하거나 분리함

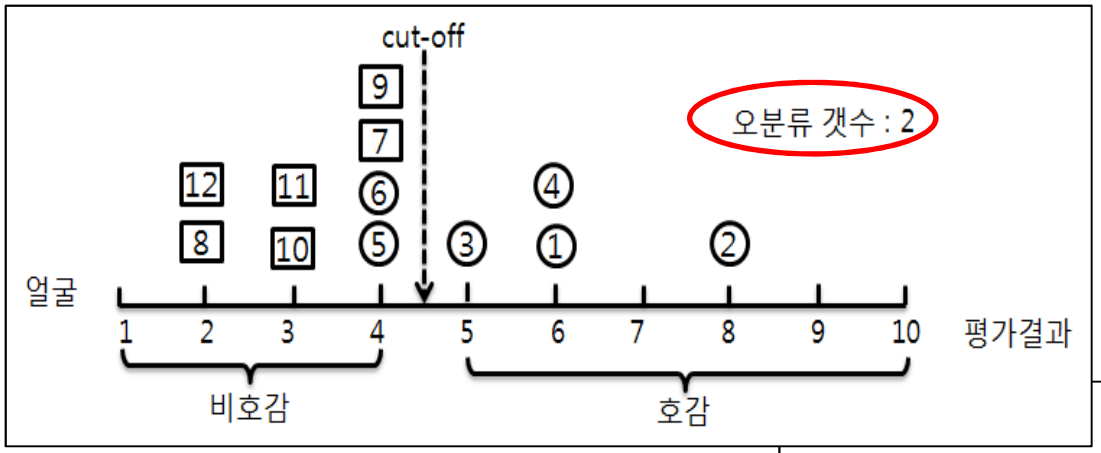
판별분석의 예 – 두 그룹 (1/4)

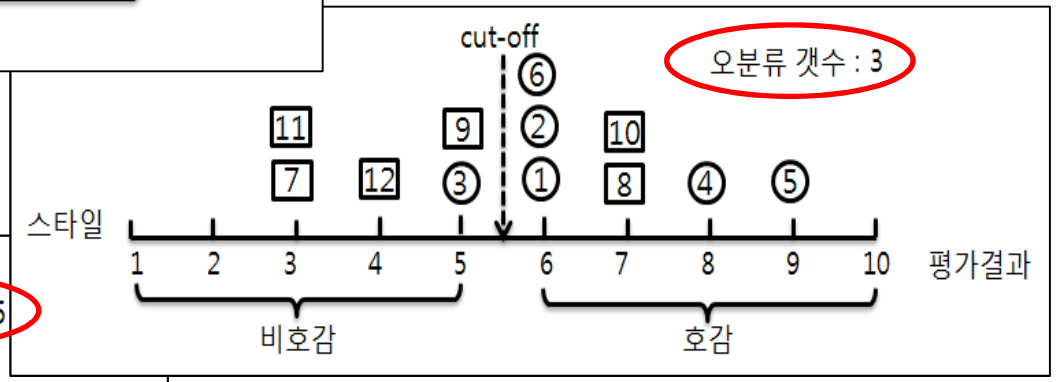
- 그룹 미팅에 대한 이성의 평가결과

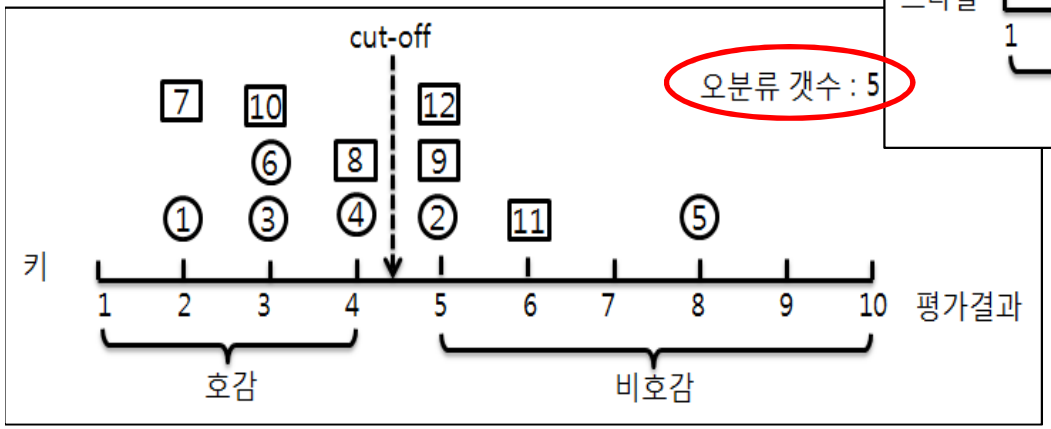
응답자와 평균	평가항목			호감여부
	얼굴 (X_1)	스타일 (X_2)	키 (X_3)	
1	6	6	2	호감
2	8	6	5	호감
3	5	5	3	호감
4	6	8	4	호감
5	4	9	8	호감
6	4	6	3	호감
그룹 평균	5.5	6.7	4.2	–
7	4	3	2	비호감
8	2	7	4	비호감
9	4	5	5	비호감
10	3	7	3	비호감
11	3	3	6	비호감
12	2	4	5	비호감
그룹 평균	3.0	4.8	4.2	–
그룹간 평균 차이	2.5	1.8	0.0	–

판별분석의 예 – 두 그룹 (2/4)

- 그룹 미팅에 대한 이성의 평가결과







판별분석의 예 – 두 그룹 (3/4)

- 그룹 미팅에 대한 이성의 평가결과

응답자와 평균	판별함수 Z		
	$Z_1=X_1$	$Z_2=X_1+X_2$	$Z_3=-5.55+0.762X_1+0.412X_2-0.14X_3$
1	6	12	1.566
2	8	14	3.048
3	5	10	0.378
4	6	14	2.362
5	4	13	1.194
6	4	10	0.028
7	4	7	-1.194
8	2	9	-1.098
9	4	9	-0.412
10	3	10	-0.322
11	3	6	-2.012
12	2	6	-2.348
cut-off	4.5	9.5	0

오분류

판별분석의 예 – 두 그룹 (4/4)

- 그룹 미팅에 대한 이성의 평가결과 – 분류정확도

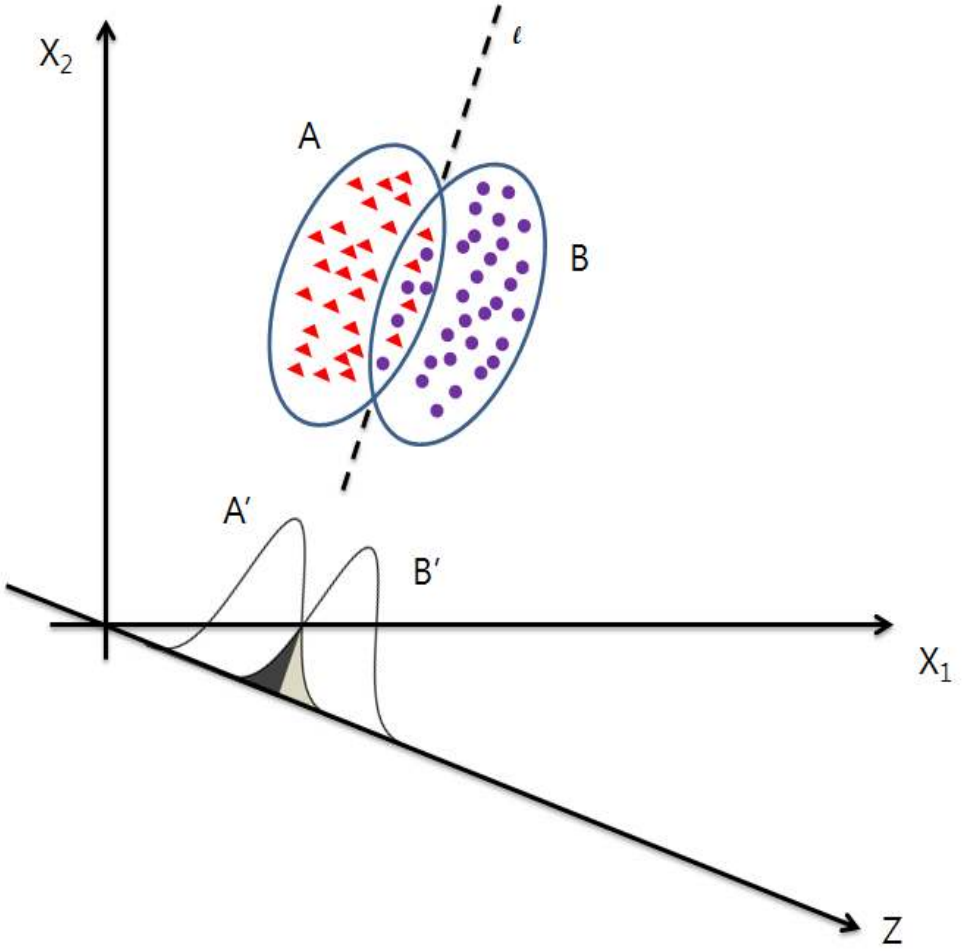
예측 / 실제	Z_1		Z_2		Z_3	
	호감	비호감	호감	비호감	호감	비호감
호감	4	2	6	0	6	0
비호감	0	6	1	5	0	6
정분류율	10/12		11/12		12/12	

- 오분류와 비용

판단 / 실제	무죄 (살인하지 않았음)	유죄 (살인했음)
무죄 (살인하지 않았음)	풀어줌 (1-α)	사형 (1종 오류, β)
유죄 (살인했음)	풀어줌 (2종 오류, α)	사형 (검정력, power, 1-β)

판별함수의 기하학적 표현

- 두 개의 IVs(X_1, X_2) 와 두 개의 범주를 갖는 DV (A그룹: ▲ , B그룹: ●) 로 이루어진 판별함수 (ℓ)



여러 그룹의 구분

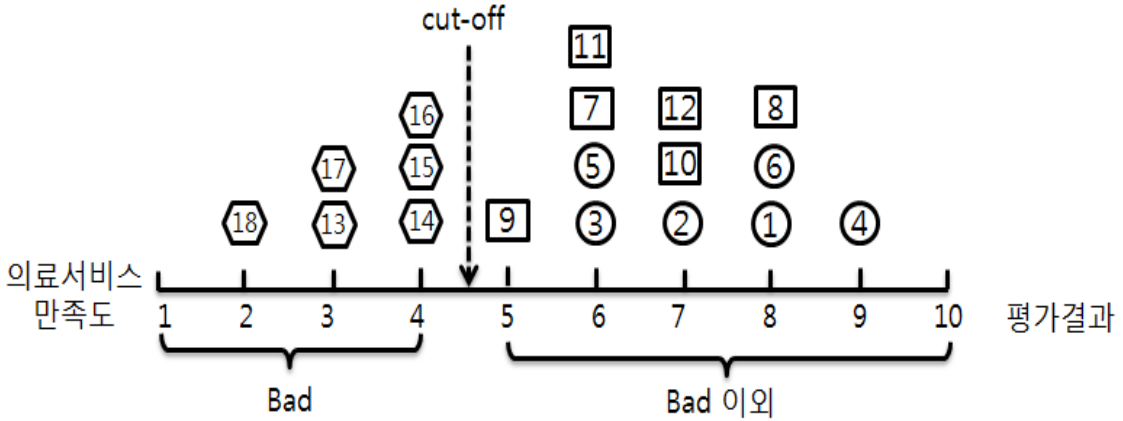
- 기본 원칙
  - $\boldsymbol{maximize}\ \frac{\text{그룹간 판별점수의 분산}}{\text{그룹내 판별점수의 분산}}$
- 판별함수의 개수
  - N개의 그룹을 분리하는 판별함수의 수는 N-1개가 사용됨

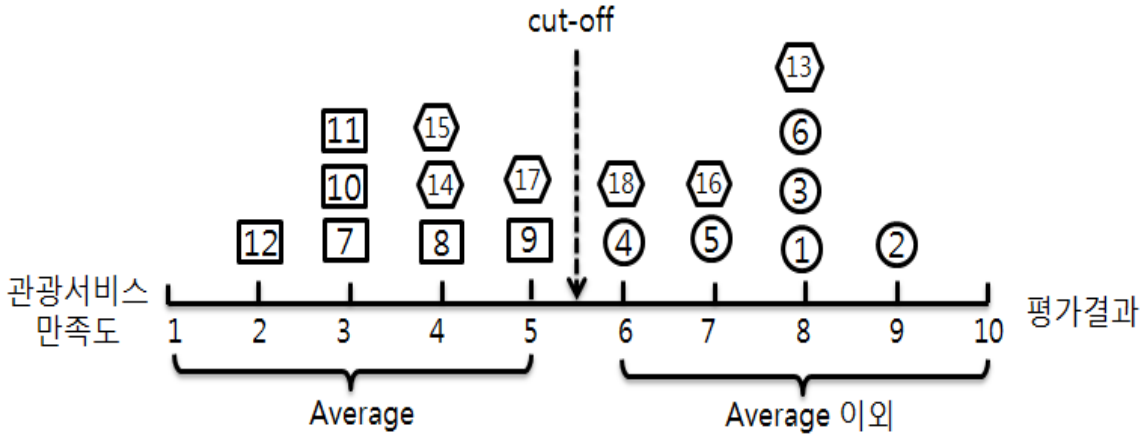
판별분석의 예 – 세 그룹(1/)

응답자와 평균	평가항목		평가결과
	의료서비스 만족도 (X_1)	관광서비스 만족도 (X_2)	
1	7	8	Good
2	7	9	Good
3	6	7	Good
4	9	6	Good
5	6	8	Good
6	8	8	Good
그룹 평균	7.2	7.5	-
7	6	3	Average
8	8	4	Average
9	5	5	Average
10	7	3	Average
11	6	4	Average
12	7	2	Average
그룹 평균	6.5	3.3	-
13	3	8	Bad
14	4	4	Bad
15	4	2	Bad
16	4	7	Bad
17	3	5	Bad
18	2	6	Bad
그룹 평균	3.3	5.7	-

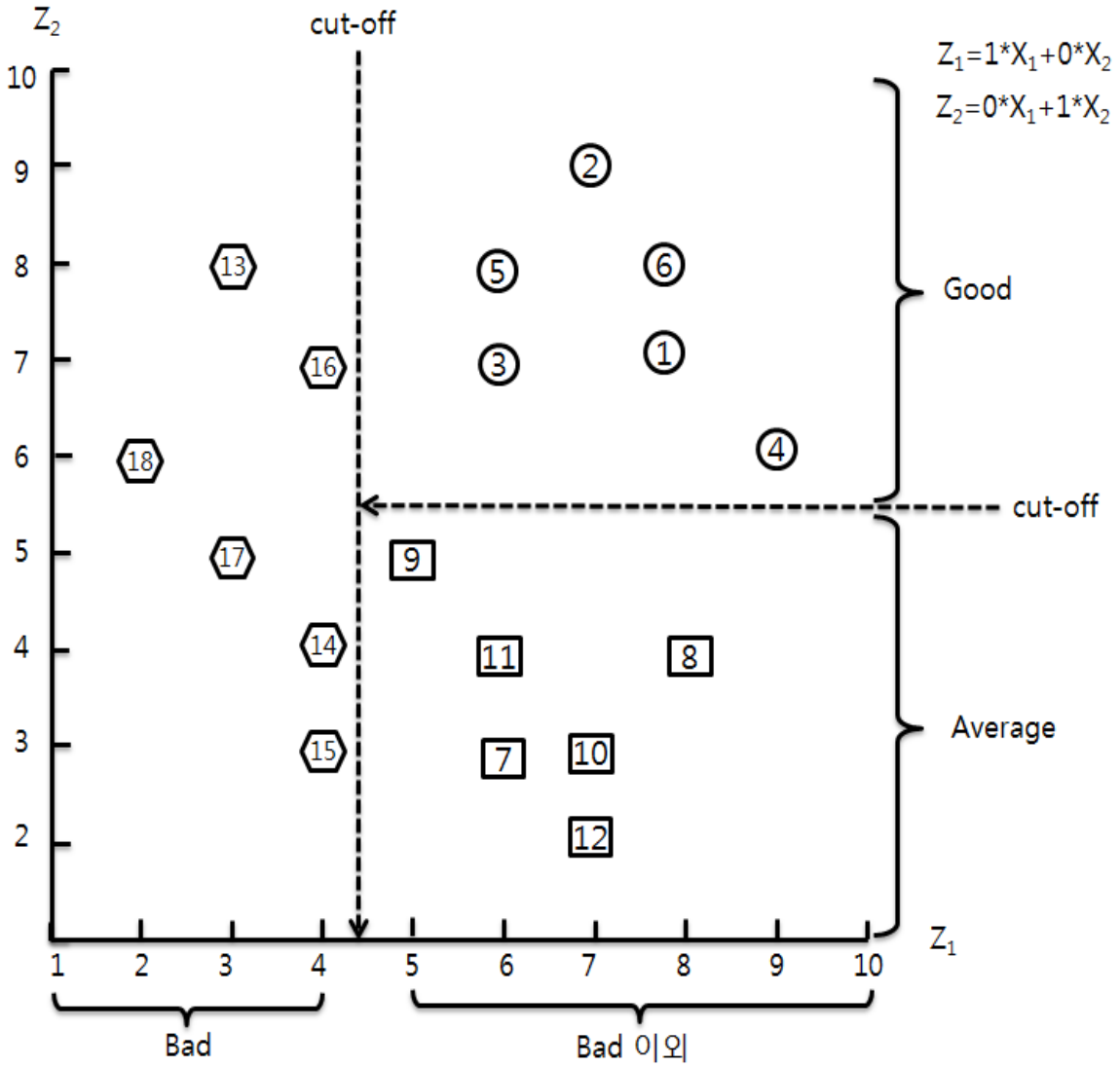
판별분석의 예 – 세 그룹(1/2)

- 원, 사각형, 육각형 안의 숫자는 평가자의 번호를 나타내며 원은 평가결과가 Good인 경우, 사각형은 평가결과가 Average인 경우, 육각형은 평가결과가 Bad인 경우를 나타냄





판별분석의 예 – 세 그룹(2/2)

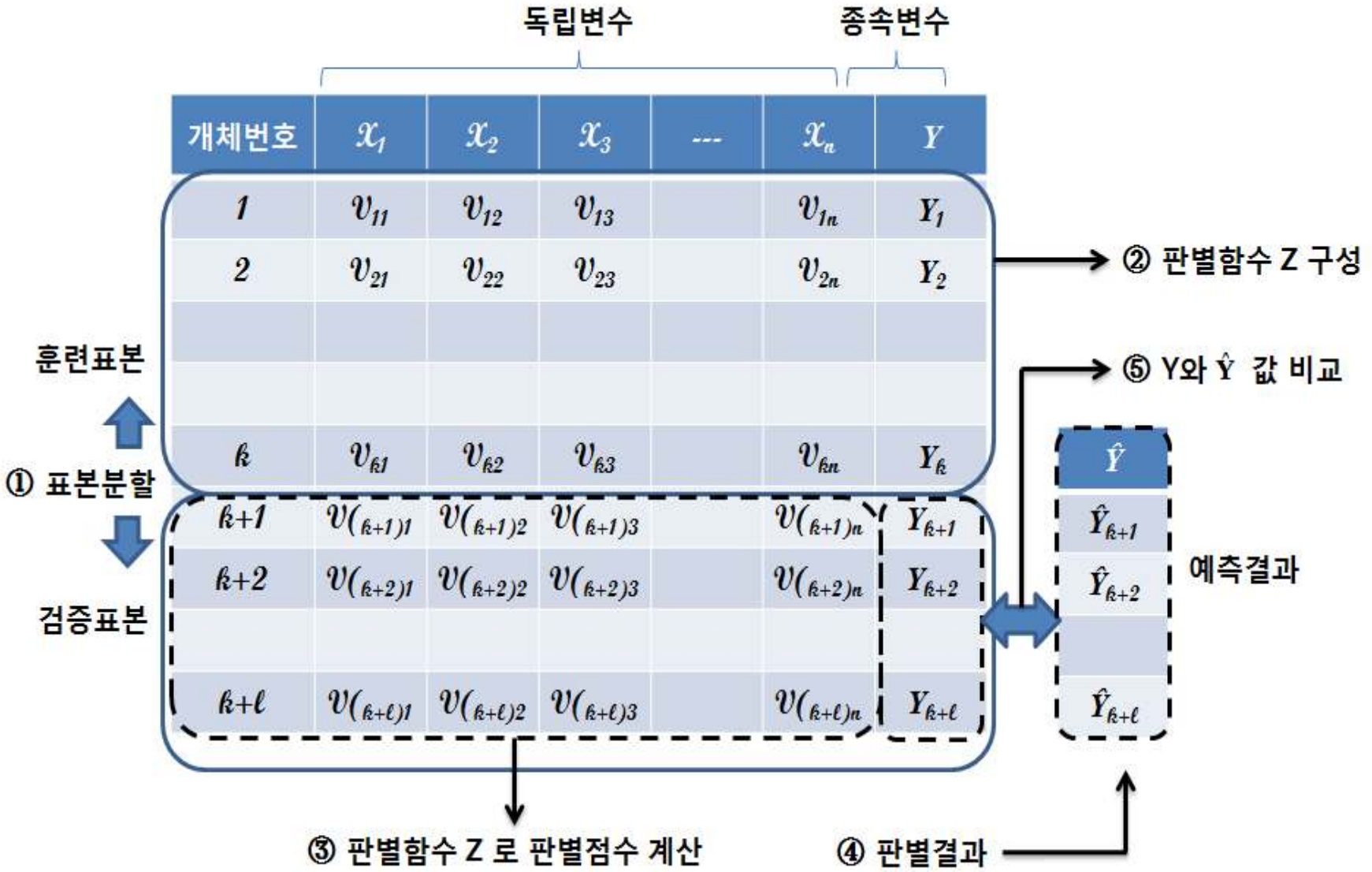


판별분석의 순서

1. 독립변수와 종속변수의 결정
2. 표본의 선정과 분할
3. 판별함수의 예측
4. 통계적 유의성 검정
5. 타당성 검증

2. 표본의 선정과 분할

- 표본의 규모
  - 독립 변수별로 20개 이상의 표본을 권장 (최소한 5개 이상) 을 필요로 한다.
  - 각 그룹별로 최소 20개 이상의 표본을 권장
  - 그룹의 사례수를 비슷한 수준으로 샘플링
- 표본의 분할
  - 표본을 분석표본(Analysis sample) 과 예비표본 (Holdout sample) 으로 분할. 나중에 판별함수를 구성하고 검증하는 교차 타당성 (cross-validation) 기법을 사용


독립변수
종속변수
개체번호
x_1
x_2
x_3
---
x_n
Y
1
v_{11}
v_{12}
v_{13}
v_{1n}
Y_1
2
v_{21}
v_{22}
v_{23}
v_{2n}
Y_2
k
v_{k1}
v_{k2}
v_{k3}
v_{kn}
Y_k
훈련표본
① 표본분할
검증표본
k+1
$v_{(k+1)1}$
$v_{(k+1)2}$
$v_{(k+1)3}$
$v_{(k+1)n}$
Y_{k+1}
k+2
$v_{(k+2)1}$
$v_{(k+2)2}$
$v_{(k+2)3}$
$v_{(k+2)n}$
Y_{k+2}
k+ℓ
$v_{(k+ℓ)1}$
$v_{(k+ℓ)2}$
$v_{(k+ℓ)3}$
$v_{(k+ℓ)n}$
$Y_{k+ℓ}$
② 판별함수 Z 구성
⑤ Y와 Ŷ 값 비교
$\hat{Y}$
$\hat{Y}_{k+1}$
$\hat{Y}_{k+2}$
$\hat{Y}_{k+ℓ}$
예측결과
③ 판별함수 Z 로 판별점수 계산
④ 판별결과

4. 통계적 유의성 검정

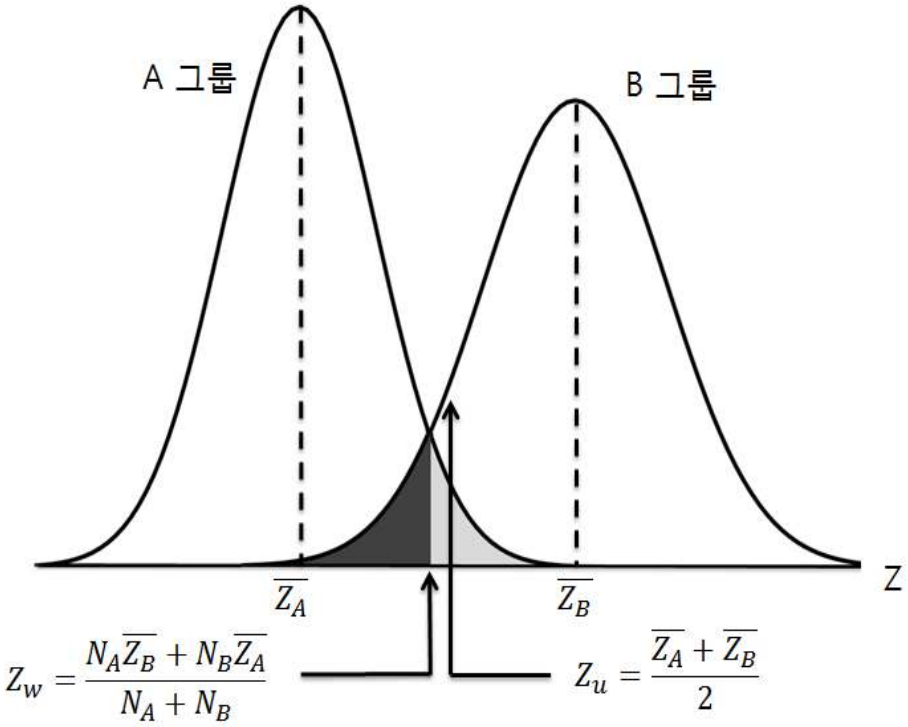
- **Wilks' lamda(Λ)**
  - **Ho: 그룹평균은 동일하다.**
  - **귀무가설을 기각할 수 없으면 판별함수가 각 그룹을 잘 구분한다고 보기 힘듦**
- **Box's M**
  - **Ho: 그룹의 공분산 행렬이 동일하다**
  - **'그룹의 공분산 행렬이 동일하다' 라는 가설이 기각되면 기본 가정을 위배하고 있으므로 표본을 늘인 후에 다시 검정하거나 비선형의 판별함수를 구하는 로짓분석 (logit analysis)과 같은 방법을 사용하는 것을 고려**

5. 타당성 검증(1/2)

- **정분류율**
  - **바르게 분류한 비율**
  - **판별점수를 나누는 cutting score를 정해야 함**
  - **정분류율을 최대화, 오분류율을 최소화하는 cutting score로 결정**
    - **최적의 기준 계산 공식**

- $Z_{cut} = \frac{N_A\overline{Z_B}+N_B\overline{Z_A}}{N_A+N_B}$,

Z_{cut}: 최적의 분리 판별점수
N_A: 그룹 A의 사례수
N_B: 그룹 B의 사례수
$\overline{Z_A}$: 그룹 A의 중심
$\overline{Z_B}$: 그룹 B의 중심



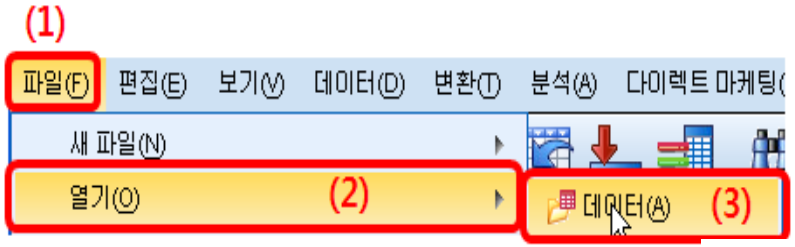
- **Student's t-test를 통해 예측정확도가 50%인지 검정**
  - $t = \frac{p-0.5}{\sqrt{\frac{0.5(1-0.5)}{N}}}$, **N은 사례수, p는 정분류율**
  - **Excel의 tdist(t, N-1) 이용**
  - **만일 샘플이 6:4라면? 0.5대신 0.6을 공식에 사용**

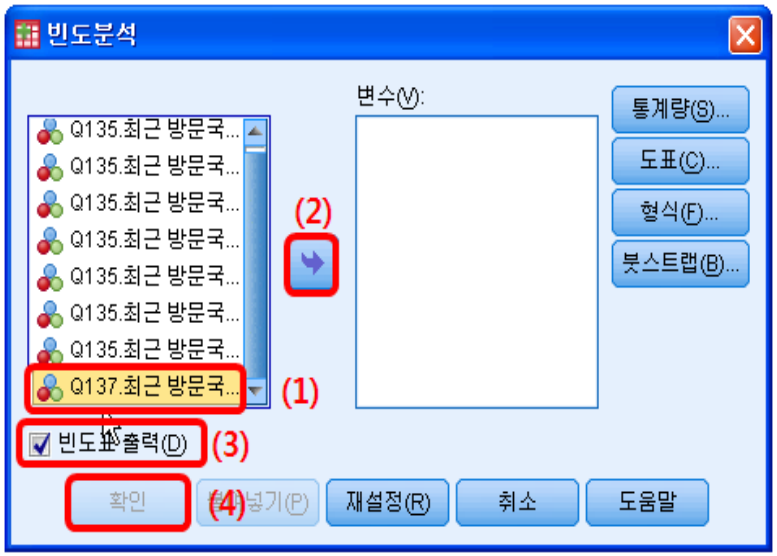
5. 타당성 검증(2/2)

- **Press's Q**
  - $\boldsymbol{Press's\ Q = \frac{\{N-(nK)\}^2}{N(K-1)}}$, **N:총 사례 수, n: 정분류 사례 수, K: 그룹의 수**
  - **통계량에 근사, χ^2통계량과 비교하여 가설 검정**
    - **Ho: 판별능력은 50%이다.**

실제 \ 예측		Z_1		Z_2		Z_3	
		호감	비호감	호감	비호감	호감	비호감
호감		4	2	6	0	6	0
비호감		0	6	1	5	0	6
정분류율		10/12		11/12		12/12	
Press's Q		$\frac{\{12-(10\times 2)\}^2}{12(2-1)}=5.3$		$\frac{\{12-(11\times 2)\}^2}{12(2-1)}=8.3$		$\frac{\{12-(12\times 2)\}^2}{12(2-1)}=12$	
검정 통계량	$\chi^2(0.01, 1)$	6.63					
	$\chi^2(0.05, 1)$	3.84					
	$\chi^2(0.1, 1)$	2.71					

SPSS 이용과 보고서 작성



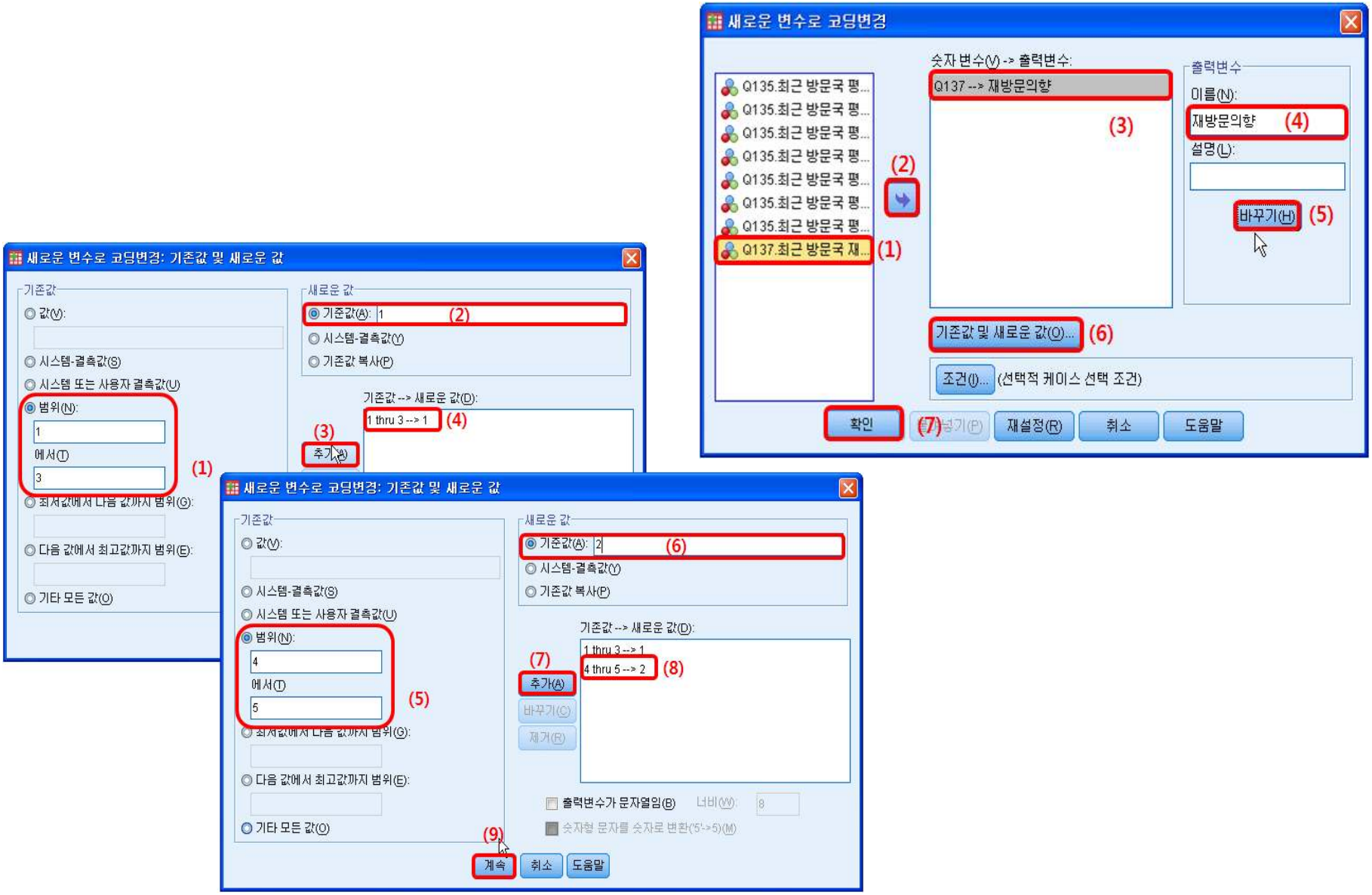


SPSS 이용과 보고서 작성

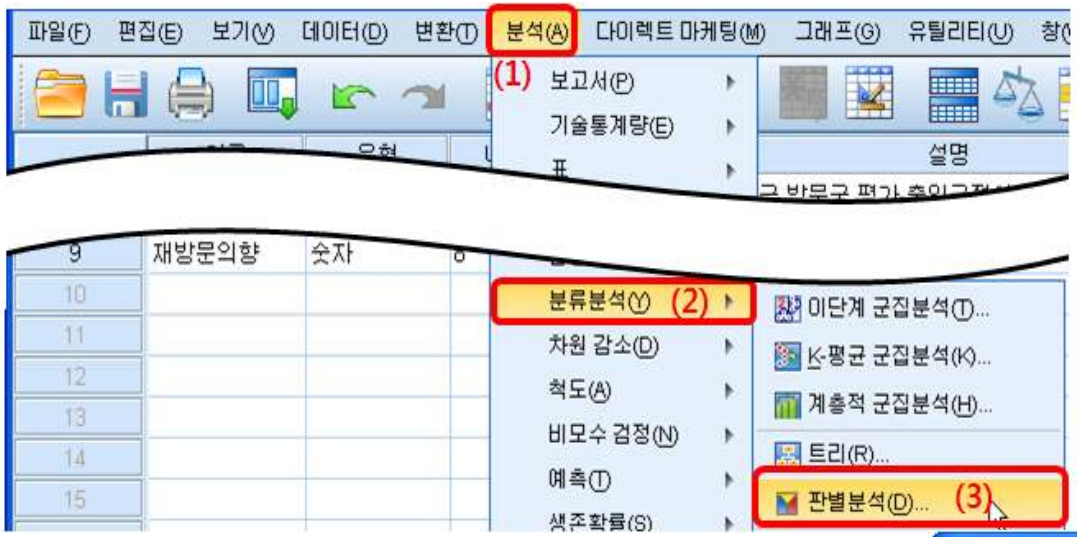
빈도가 너무 작음, 재코딩이 필요

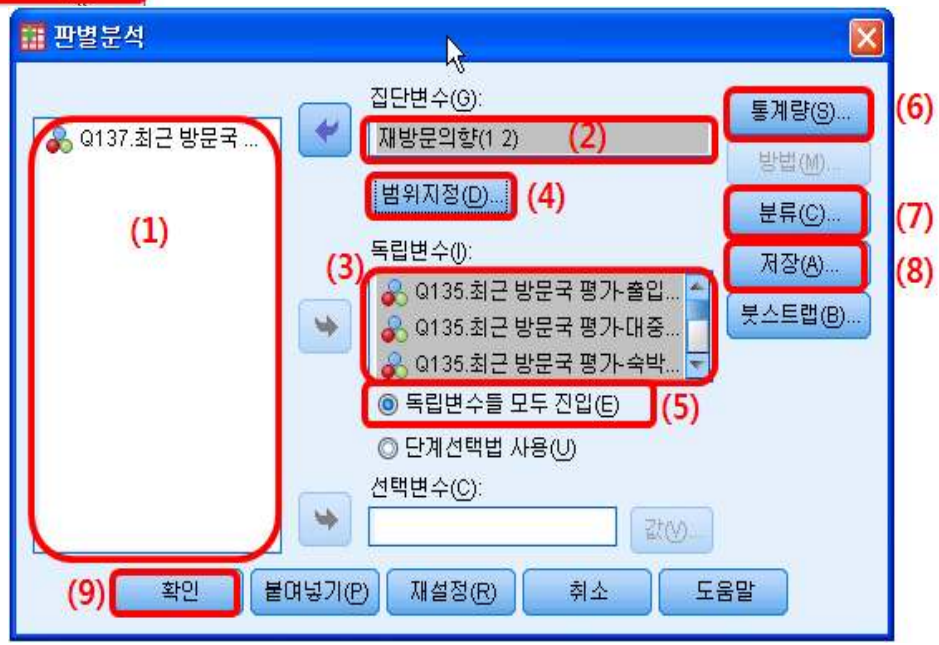
		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	전혀 그렇지 않다	1	.1	.3	.3
	별로 그렇지 않다	6	.6	1.9	2.3
	보통이다	49	5.3	15.8	18.1
	약간 그런 편이다	153	16.5	49.4	67.4
	매우 그렇다	101	10.9	32.6	100.0
결측	시스템 결측값	616	66.5		
합계		926	100.0	100.0	

SPSS 이용과 보고서 작성

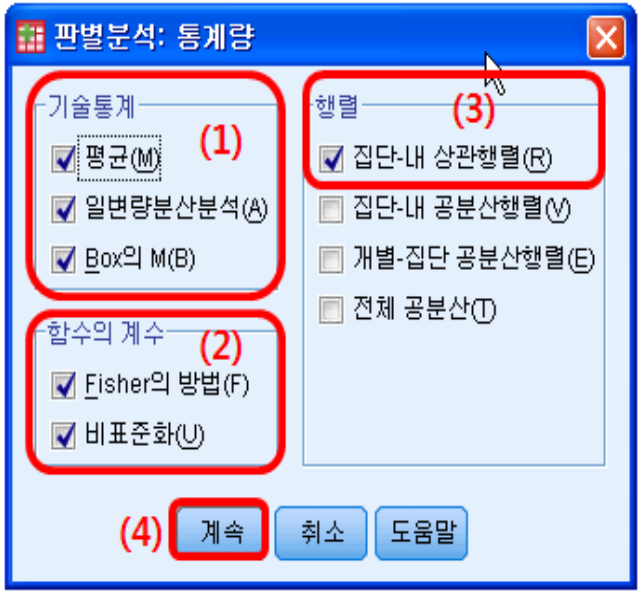


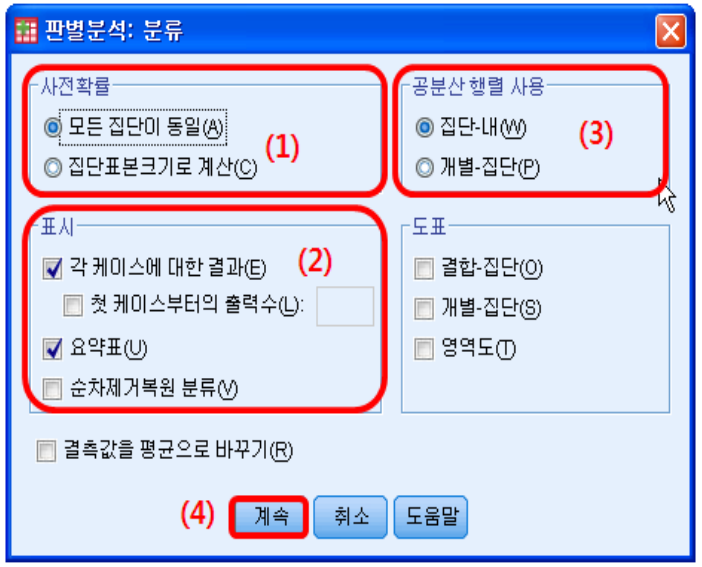
SPSS 이용과 보고서 작성

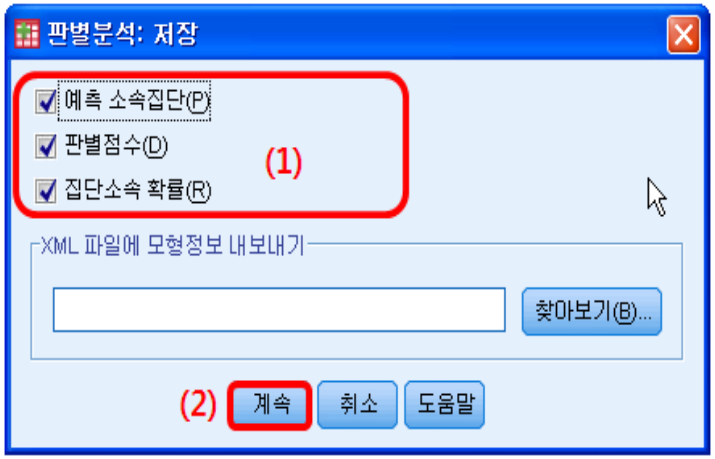




SPSS 이용과 보고서 작성







결과

변수명	분류	변수설명	척 도
재방문의향	종속변수	낮음 = 1, 높음 = 2	명목척도
출입국 절차 만족도	독립변수	매우 불만족 = 1 불만족 = 2 보통 = 3 만족 = 4 매우 만족 = 5	등간척도
대중교통 만족도	독립변수		등간척도
숙박	독립변수		등간척도
음식	독립변수		등간척도
쇼핑	독립변수		등간척도
관광지 매력도	독립변수		등간척도
관광정보 입수 용이성	독립변수		등간척도

- **Box's M**
  - **Ho: 그룹별 공분산 행렬이 동일하다.**

Box의 M		32.043
F	근사법	1.090
	자유도1	28
	자유도2	34682.112
	유의확률	.338

결과

- **Wilks' lamda(Λ)**
  - **Ho: 그룹간 평균이 동일하다.**

	Wilks 람다	F	자유도1	자유도2	유의확률
출입국 절차 만족도	.962	12.034	1	308	.001
대중교통 만족도	.912	29.883	1	308	.000
숙박	.878	42.626	1	308	.000
음식	.907	31.422	1	308	.000
쇼핑	.910	30.351	1	308	.000
관광지 매력도	.896	35.905	1	308	.000
관광정보 입수 용이성	.915	28.477	1	308	.000

- **판별함수의 Wilks' lamda(Λ)**
  - **Ho: 판별함수에 이용 유무에 따른 판별에 차이가 없다.**

Wilks' lamda	카이제곱	자유도	유의확률
.841	52.843	7	.000

결과

- 판별함수의 표준화 비표준화 계수

평가항목 (독립변수)	표준화 계수	비표준화 계수
출입국절차	-.307	-.335
대중교통	.327	.380
숙박	.490	.607
음식	.207	.250
쇼핑	.074	.095
관광지매력도	.270	.340
관광정보 입수 용이성	.101	.137
(상수)		-5.659

- 판별함수는
  - **Z=-5.659-0.335×출입국절차+0.380×대중교통+0.607×숙박+0.250×음식+0.095×쇼핑+0.340×관광지매력도+0.137×관광정보 입수 용이성**

결과

- 분류정확도

재방문의향		예측 그룹		전체
		1.00	2.00	
원래 그룹	1.00	42	14	56
	2.00	53	201	254
계		95	215	310

  - 분류정확도를 나타내는 정분류율은 (42+201)/310 = 78.4%

- Student's t
  - $t = \frac{0.784-0.5}{\sqrt{\frac{0.5(1-0.5)}{310}}} = 10.0 > Z_{\frac{0.05}{2}} = 1.96$
- Press's Q
  - $Press's\ Q = \frac{\{N-(nK)\}^2}{N(K-1)} = \frac{\{310-243\times2\}^2}{310(2-1)} = 99.9 > 3.84 = \chi^2(0.05,\ 1)$