

## 응용통계학및실습 과제2

20181256 수학과 김도현

9-4.

(가)

해당 분할표는 회사 직급에 따른 연금계획에 대한 표를 나타낸다. 아마 직급과 연금계획 두 변수 간의 연관성에 대해 파악하고자 하는 표로 분석해볼 수 있다. 미리 일정하게 부모집단을 뽑아 놓고 시작했기 때문에 동일성 검정이라 유추 할 수 있다.

(나)

H0: 귀무가설은 직급에 따른 연금계획에 대한 두 변수는 서로 동질하다 라고 세운다.

H1: 대립가설은 직급에 따른 연금계획에 대한 두 변수는 서로 동질하지 않다 라고 세운다.

(다)

```
1 data company;
2 input level plan $ count @@;
3 cards;
4 1 A 140 1 B 40 1 C 20
5 2 A 70 2 B 20 2 C 10
6 3 A 160 3 B 30 3 C 10
7 4 A 80 4 B 10 4 C 10
8 ;
9 run;
L0 proc freq data=company order=data;
L1 weight count;
L2 tables level*plan / chisq expected cellchi2 measures;
L3 run;
L4
```

level \* plan 테이블에 대한 통계량

통계량	자유도	값	Prob
카이제곱	6	11.0000	0.0884
우도비 카이제곱	6	11.7111	0.0687
Mantel-Haenszel 카이제곱	1	4.7064	0.0301
파이 계수		0.1354	
우발성 계수		0.1342	
크래머의 V		0.0957	

해당 통계량을 보면 카이제곱 Prob 가 0.0884 가 나왔음을 확인할 수 있다. 이는 H0를 약 9 퍼센

트 유의 수준에서 기각 할 수 있음을 알 수 있다.

(라)

**FREQ 프로시저**

빈도 기대값 셀 카이제곱 백분율 행 백분율 칼럼 백분율	테이블 level * plan				
	level	plan			합계
		A	B	C	
<b>1</b>		140	40	20	200
		150	33.333	16.667	
		0.6667	1.3333	0.6667	
		23.33	6.67	3.33	33.33
		70.00	20.00	10.00	
		31.11	40.00	40.00	
<b>2</b>		70	20	10	100
		75	16.667	8.3333	
		0.3333	0.6667	0.3333	
		11.67	3.33	1.67	16.67
		70.00	20.00	10.00	
		15.56	20.00	20.00	
<b>3</b>		160	30	10	200
		150	33.333	16.667	
		0.6667	0.3333	2.6667	
		26.67	5.00	1.67	33.33
		80.00	15.00	5.00	
		35.56	30.00	20.00	
<b>4</b>		80	10	10	100
		75	16.667	8.3333	
		0.3333	2.6667	0.3333	
		13.33	1.67	1.67	16.67
		80.00	10.00	10.00	
		17.78	10.00	20.00	
<b>합계</b>		450	100	50	600
		75.00	16.67	8.33	100.00

(Executive, A) 는 우리 데이터에서는 (2,A) 를 의미한다. 이때 셀 카이제곱은 0.3333 이 나왔다. 다른 것들에 비해 셀 카이제곱 값이 그렇게 크지 않음을 확인할 수 있다. 그리고 빈도는 70, 기대값은 75가 나왔다. 이 역시 큰 차이를 보이지 않으니 두 변수 간의 연관관계에 공헌하는 정도가 그렇게 크지 않다고 판단할 수 있다.

(마)

통계량	값	ASE
감마	-0.1673	0.0678
Kendall의 타우-b	-0.0906	0.0367
Stuart의 타우-c	-0.0733	0.0298
Somers D CIR	-0.0677	0.0276
Somers D RIC	-0.1214	0.0491
Pearson 상관계수	-0.0886	0.0418
Spearman 상관계수	-0.1006	0.0408
람다 비대칭 CIR	0.0000	0.0000
람다 비대칭 RIC	0.0500	0.0422
람다 대칭	0.0364	0.0309
불확실 계수 CIR	0.0135	0.0076
불확실 계수 RIC	0.0073	0.0041
불확실 계수 대칭	0.0095	0.0053

표본 크기 = 600

두 변수 간의 유의관계를 파악하기 위해 measures option을 썼다. 우선 크래머의 v 값은 1과 매우

멀다. 또한 measures option을 통한 값들도 대부분 작기 때문에 그렇게 큰 연관관계를 가진다고는 할 수 없다. 하지만 연관성 측도/ASE 의 절대값을 확인해보면 2를 대부분 넘는다. 즉 이는 두 변수에 대한 연관성이 0이다를 기각한다고 볼 수 있다.

9-6.

```
data vegetable;
input kind atmosphere count @@;
cards;
1 0 32 1 1 8
2 0 28 2 1 12
3 0 19 3 1 21
;
run;
proc freq data=vegetable order=data;
weight count;
tables kind*atmosphere / chisq;
run;
```

FREQ 프로시저

빈도 백분율 행 백분율 칼럼 백분율	테이블 kind * atmosphere			
	kind	atmosphere		합계
		0	1	
1		32	8	40
		26.67	6.67	33.33
		80.00	20.00	
		40.51	19.51	
2		28	12	40
		23.33	10.00	33.33
		70.00	30.00	
		35.44	29.27	
3		19	21	40
		15.83	17.50	33.33
		47.50	52.50	
		24.05	51.22	
합계		79	41	120
		65.83	34.17	100.00

kind \* atmosphere 테이블에 대한 통계량

통계량	자유도	값	Prob
카이제곱	2	9.8549	0.0072
우도비 카이제곱	2	9.8593	0.0072
Mantel-Haenszel 카이제곱	1	9.3135	0.0023
파이 계수		0.2866	
우발성 계수		0.2755	
크래머의 V		0.2866	

표본 크기 = 120

우선 각 야채 종류 별 데이터 개수 40개로 같기 때문에 미리 처음부터 구분해서 뽑았다 판단한다. 이는 동일성 검정으로 판단할 수 있다.

H0: 야채 종류에 따라 대기오염이 미치는 영향이 동일하다.

H1: 야채 종류에 따라 대기오염이 미치는 영향이 동일하지 않다.

통계량의 카이 제곱을 보면 0.0072 임을 확인할 수 있다. 이는 약 1퍼센트 유의수준으로 H0를 기

각한다는 말이다. 우리 과제에서는 5퍼센트 유의수준에서 검정하라고 했기 때문에  $H_0$ 를 기각할 수 있다. 즉 야채 종류에 따라 대기오염이 미치는 영향이 동일하지 않음을 알 수 있다.

9-9.

```
data money;
input place level count @@;
cards;
1 1 76 1 2 53 1 3 59 1 4 48
2 1 124 2 2 147 2 3 141 2 4 152
;
run;
proc freq data=money order=data;
weight count;
tables place*level / chisq measures;
run;
```

## FREQ 프로시저

빈도  
백분율  
행 백분율  
칼럼 백분율

		테이블 place * level				
place		level				합계
		1	2	3	4	
1		76	53	59	48	236
		9.50	6.63	7.38	6.00	29.50
		32.20	22.46	25.00	20.34	
		38.00	26.50	29.50	24.00	
2		124	147	141	152	564
		15.50	18.38	17.63	19.00	70.50
		21.99	26.06	25.00	26.95	
		62.00	73.50	70.50	76.00	
합계		200	200	200	200	800
		25.00	25.00	25.00	25.00	100.00

### place \* level 테이블에 대한 통계량

통계량	자유도	값	Prob
카이제곱	3	10.7224	0.0133
우도비 카이제곱	3	10.5353	0.0145
Mantel-Haenszel 카이제곱	1	7.3042	0.0069
파이 계수		0.1158	
우발성 계수		0.1150	
크래머의 V		0.1158	

표본 크기 = 800

무작위로 도시와 농촌에서 800명의 표본을 무작위로 추출했다고 가정하였다. 이는 독립성 검정으로 파악할 수 있다.

$H_0$ : 지역과 소득 수준은 서로 독립이다.

$H_1$ : 지역과 소득 수준은 서로 독립이 아니다.

통계량의 카이제곱 Prob 값을 보면 0.0133이 나왔다. 이는 약 2퍼센트 유의수준에서  $H_0$ 가 기각할 수 있음을 알 수 있다. 그래서 과제에서 요구한 5퍼센트 유의수준 이내에 들어오기 때문에  $H_0$ 를 기각한다. 즉 지역과 소득 수준은 서로 독립이 아님을 알 수 있다. 조금 더 자세히 보기 위해 measures option도 추가했다.

파이 계수		0.1158	
우발성 계수		0.1150	
크레머의 V		0.1158	

통계량	값	ASE
감마	0.1556	0.0572
Kendall의 타우-b	0.0873	0.0323
Stuart의 타우-c	0.0975	0.0362
Somers D CIR	0.1172	0.0434
Somers D RIC	0.0650	0.0242
Pearson 상관계수	0.0956	0.0354
Spearman 상관계수	0.0956	0.0354
람다 비대칭 CIR	0.0467	0.0270
람다 비대칭 RIC	0.0000	0.0000
람다 대칭	0.0335	0.0195
불확실 계수 CIR	0.0047	0.0029
불확실 계수 RIC	0.0109	0.0067
불확실 계수 대칭	0.0066	0.0041

표본 크기 = 800

해당 measures 옵션문을 쓰면 이렇게 나오는데 해당 값들은 1과는 조금 거리가 있어 약한 연관성이 있다. 일반적으로 연관성 측도/ASE 가 약 2 이상이기 때문에 5% 유의수준 내에서 두 변수 간의 연관성이 0이 아님을 알 수 있다.