**SAS 데이터 분석 입문 9장**

**2019020650 김형욱**

**\* 9장 예제문제**

**<예 9-1>**

**data** sasadv.soft;

input age $ beverage $ count @@;

cards;

20대 coke 10 20대 pepsi 14 20대 fanta 4 20대 others 12

30대 coke 13 30대 pepsi 9 20대 fanta 10 30대 others 8

40대 coke 12 40대 pepsi 8 20대 fanta 10 40대 others 10

;

**run**;

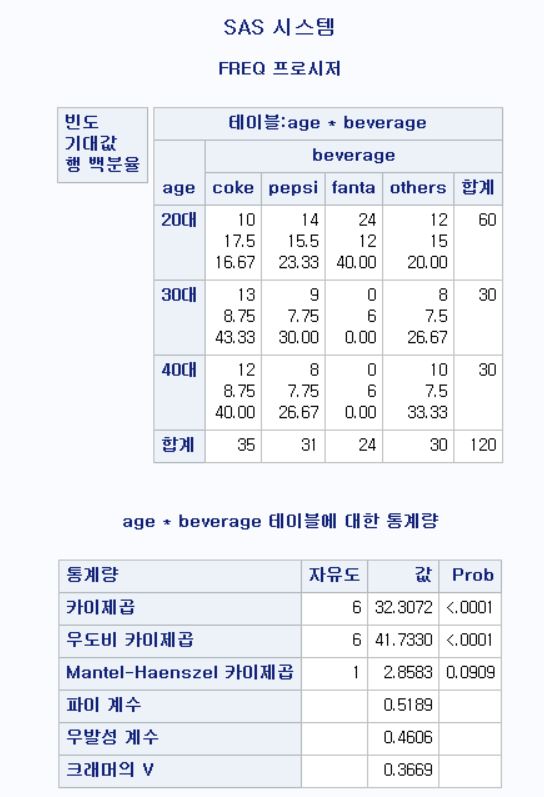
**proc** **freq** data=sasadv.soft order=data;

weight count;

exact fisher;

tables age\*beverage / nocol nopercent expected chisq;

**run**;



해석 : 위 문제에서 귀무가설은 “연령대에 따라 청량음료 선호도의 분포가 같다”이고, 대립가설은 “연령대에 따라 청량음료 선호도의 분포가 다르다”이다. 카이제곱 검정 결과, 카이제곱 통계량 값이 32.3072이고 우도비 카이제곱 통계량 값은 41.7330이며, 유의확률은 모두 0.0001보다 작으므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 연령대에 따라 청량음료 선호도의 분포가 다르다고 할 수 있다. 또한 표본의 수가 적기 때문에 Fisher’s exact test를 통해 검정을 해 본 결과, 마찬가지로 유의확률이 0.0001보다 작으므로 유의수준 5% 하에서 귀무가설을 기각한다.

**<예 9-2>**

**data** sasadv.bean;

input type count @@;

cards;

1 315 2 108 3 101 4 32

;

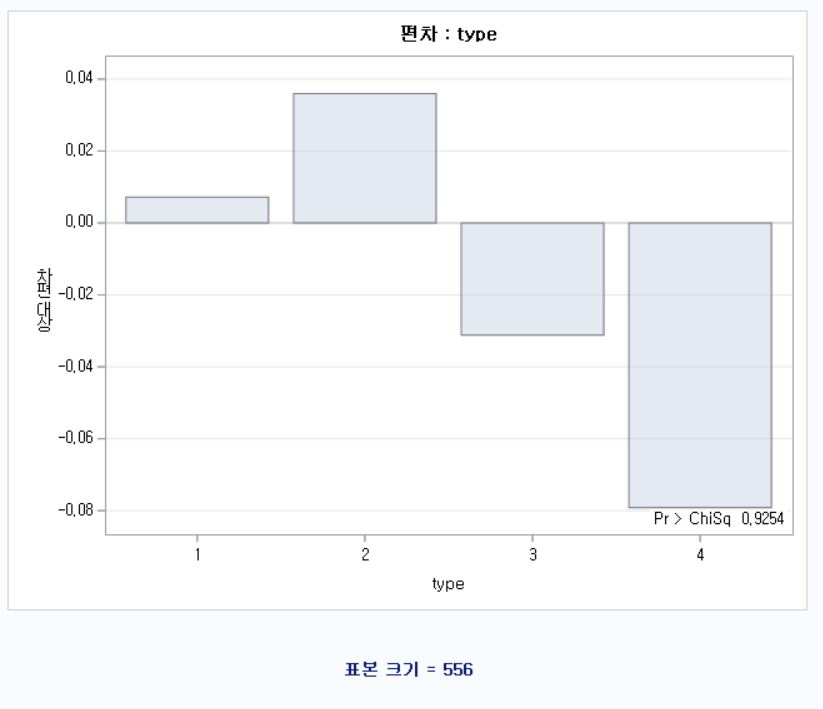
**run**;

**proc** **freq** data=sasadv.bean;

weight count;

tables type / nocum testp=(**0.5625** **0.1875** **0.1875** **0.0625**);

**run**;



해석 : 위 문제에서 귀무가설은 “멘델의 법칙에 따라 네 종류의 완두콩이 나타날 비가 9:3:3:1이다(즉, 관측된 완두콩들의 비율은 멘델의 법칙을 따른다)”이고, 대립가설은 “멘델의 법칙과 다르게 네 종류의 완두콩이 나타날 비가 9:3:3:1이 아니다(즉, 관측된 완두콩들의 비율은 멘델의 법칙을 따르지 않는다)”이다. 카이제곱 검정 결과, 카이제곱 통계량 값이 0.47이고 유의확률이 0.9254이므로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 멘델의 법칙에 따라 네 종류의 완두콩이 나타날 비가 9:3:3:1이라고 할 수 있다.

**<예 9-3>**

**data** sasadv.edueco;

input edu eco count @@;

cards;

1 1 255 1 2 105 1 3 81

2 1 110 2 2 92 2 3 66

3 1 90 3 2 113 3 3 88

;

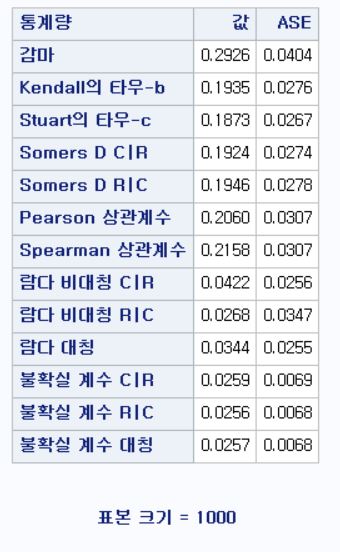
**run**;

**proc** **freq** data=sasadv.edueco order=data;

weight count;

tables edu\*eco / nocol nopercent chisq measures;

**run**;



해석 : 위 문제에서 귀무가설은 “변수 edu와 eco간에는 연관성이 없다”이고, 대립가설은 “변수 edu와 eco간에는 연관성이 있다”이다. 먼저 두 변수 사이에 연관성이 존재하는지에 대한 카이제곱 검정 결과를 살펴보면, 카이제곱 통계량 값이 54.2539이고 우도비 카이제곱 통계량 값이 54.9719이며 유의확률이 모두 0.0001보다 작으므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 변수 edu와 eco 사이에는 연관성이 존재한다고 할 수 있다. 여러 가지 연관성의 측도들을 통해 두 변수들 사이의 연관성 여부를 확인할 수가 있다. 우선 순서형 변수들 간의 연관성 측도로는 ‘감마(Gamma)’, ‘켄달의 타우-b’, ‘스튜어트의 타우-c’, ‘소머즈의 D’, ‘피어슨 상관계수’, ‘스피어만 상관계수’ 등이 있다. 그리고 명목형 변수들 간의 연관성 측도로는 ‘비대칭 람다(C|R)’, ‘비대칭 람다(R|C)’, ‘대칭 람다’, ‘엔트로피에 기초한 불확실성 계수’ 등이 있다.

**\* 9장 연습문제**

**<연습문제 9-3>**

**data** sasadv.ex9\_3;

input region $ candidate $ count @@;

cards;

서울 노태우 1683 서울 김영삼 1637 서울 김대중 1833 서울 김종필 461

부산 노태우 641 부산 김영삼 1117 부산 김대중 182 부산 김종필 52

대구 노태우 800 대구 김영삼 275 대구 김대중 30 대구 김종필 23

인천 노태우 326 인천 김영삼 249 인천 김대중 177 인천 김종필 76

광주 노태우 23 광주 김영삼 2 광주 김대중 450 광주 김종필 1

;

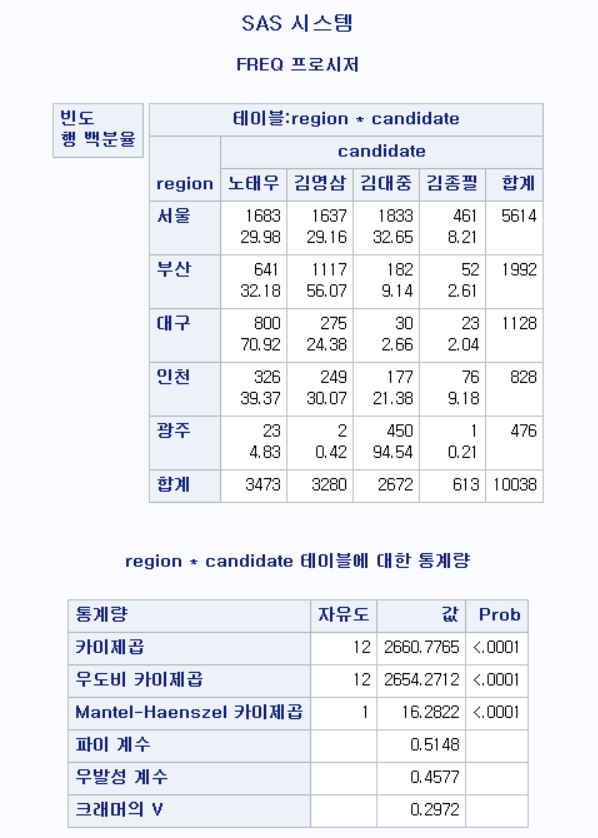
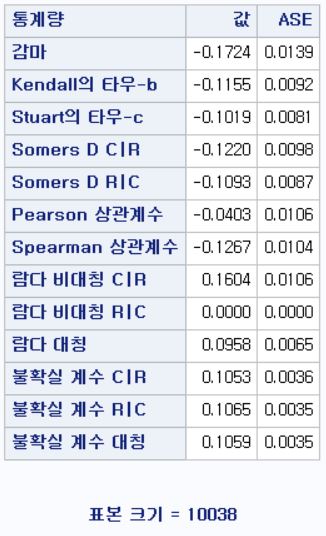
**run**;

**proc** **freq** data=sasadv.ex9\_3 order=data;

weight count;

tables region\*candidate / nocol nopercent chisq measures;

**run**;

해석 : 위 문제에서 귀무가설은 “지역과 후보자에 대한 지지 간에 관계가 없다”이고, 대립가설은 “지역과 후보자에 대한 지지 간에 관계가 있다”이다. 먼저 카이제곱 검정 결과를 살펴보면, 카이제곱 통계량 값이 2660.7765이고 유의확률이 0.0001보다 작으므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 지역과 후보자에 대한 지지 간에 관계가 있다고 할 수 있다. 또한 변수 region과 candidate 모두 명목형 변수이므로, 명목형 변수들 간의 연관성 측도를 나타내주는 ‘비대칭 람다(C|R)’, ‘비대칭 람다(R|C)’, ‘대칭 람다’, ‘엔트로피에 기초한 불확실성 계수’ 등을 통해 연관성의 정도를 파악할 수도 있다.

**<연습문제 9-5>**

**data** sasadv.ex9\_5;

input color $ count @@;

cards;

흰색 69 상아색 75 올리브색 67 회색 49

;

**run**;

**proc** **freq** data=sasadv.ex9\_5;

weight count;

tables color / nocum testp=(**0.2654** **0.2885** **0.2577** **0.1885**);

**run**;



해석 : 위 문제에서 귀무가설은 “세탁기 색상에 따라 주부들의 선호도에 차이가 없다”이고, 대립가설은 “세탁기 색상에 따라 주부들의 선호도에 차이가 있다”이다. 카이제곱 적합도 검정 결과, 카이제곱 검정통계량 값이 14.3666이고 유의확률이 0.0024이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 세탁기 색상에 따라 주부들의 선호도에 차이가 있다고 할 수 있다.

**<연습문제 9-9>**

**data** sasadv.ex9\_9;

input region $ income $ count @@;

cards;

도시 상 76 도시 중상 53 도시 중하 59 도시 하 48

농촌 상 124 농촌 중상 147 농촌 중하 141 농촌 하 152

;

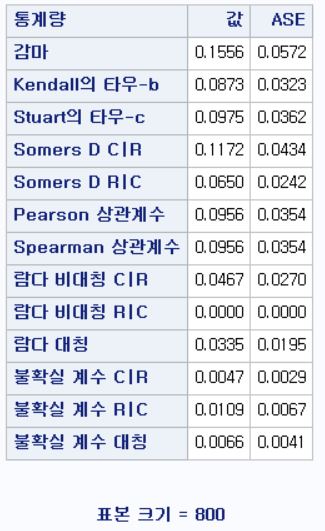
**run**;

**proc** **freq** data=sasadv.ex9\_9 order=data;

weight count;

tables region\*income / nocol nopercent chisq measures;

**run**;



해석 : 위 문제에서 귀무가설은 “지역에 따른 소득수준에 차이가 존재하지 않는다”이고, 대립가설은 “지역에 따른 소득수준에 차이가 존재한다”이다. 카이제곱 검정 결과, 카이제곱 검정통계량 값이 10.7224이고 유의확률이 0.0133이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 지역에 따른 소득수준에 차이가 존재한다고 할 수 있다.