경영데이터분석기초

- SPSS, Excel을 활용한 통계분석 -

유 진 호

jhyoo@smu.ac.kr

지난 Summary

- 데이터 불러오기
- 평균, 편차, 제곱합, 분산, 표준편차, 최빈값, 왜도, 첨도
- 중위수, Q1/Q2/Q3, Box-Plot, Stem-Leaf
- 정규성 검증(Q-Q Plot, K-S, S-W 통계량)
- 공분산, 상관계수

척도와 분석간의 관계		독립변수 X	
		범주형 자료	연속형 자료
종속변수 Y	범주형 자료	교차분석 (x^2검정)	로지스틱 회귀분석 판별분석
	연속형 자료	ANOVA(분산분석)	회귀분석

교차분석(카이제곱 검정)

연구자가 복잡한 자료를 상황표로 만들어서, 변수 사이의 상관관계를 파악할 수 있는 것이 교차분석이다. 교차분석에서 두 변수가 상호 독립적인지 아니면 관련성(연관성)이 있는지를 분석하는 것이 카이제곱 검정이다.

예를 들어 결혼상태에 따라 교육수준에 대한 차이가 있는지를 알아보기 위한 카이제곱 검정의 가설은 다음과 같다.

귀무가설 : 결혼상태에 따라 교육수준에 차이가 없다.

대립가설 : 결혼상태에 따라 교육수준에 차이가 있다.

여기서 귀무가설은 두 변수간의 관계가 독립적이라는 의미이고, 대립가설은 두 변수간의 관계가 독립적이지 않고 어떤 관계가 있음을 의미한다. 카이제곱 검정은 단지 두 변수가 독립적인지 아닌지만 알 수 있을 뿐 구체적으로 어떤 관계가 있는지는 알 수 없다.

빈도표

marital

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	divorced	528	11.7	11.7	11.7
	married	2797	61.9	61.9	73.5
	single	1196	26.5	26.5	100.0
	합계	4521	100.0	100.0	

education

		빈도	퍼센트	유효 퍼센트	누적퍼센트
유효	primary	678	15.0	15.0	15.0
	secondary	2306	51.0	51.0	66.0
	tertiary	1350	29.9	29.9	95.9
	unknown	187	4.1	4.1	100.0
	합계	4521	100.0	100.0	

marital * education 교차표

빈도

		education				전체
		primary	secondary	tertiary	unknown	
marital	divorced	79	270	155	24	528
	married	526	1427	727	117	2797
	single	73	609	468	46	1196
전체		678	2306	1350	187	4521

교차분석(카이제곱 검정)

◆ 카이제곱 통계량 계산

카이제곱 통계량은 실제의 자료에서 얻은 관찰빈도와 기대빈도의 차이를 비교함으로써, 즉 주어진 관찰빈도가 기대빈도에 얼마나 가까운지를 봄으로써 귀무가설을 검증하게 된다.

$$\chi^2 = \Sigma\Sigma \frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}}$$

통계량과 자유도와 유의수준을 이용하여 카이제곱 분포표를 찾아 카이제곱 임계값을 가지고 비교하여 카이제곱 통계량이 임계값보다 크면, 두 변수가 독립이라는 귀무가설을 기각하고, 카이제곱 통계량이 임계값보다 작으면 귀무가설을 채택한다. 귀무가설을 채택한다는 것은 두 집단간에 차이가 없다는 의미이다. 즉, 독립적이다는 의미이다.

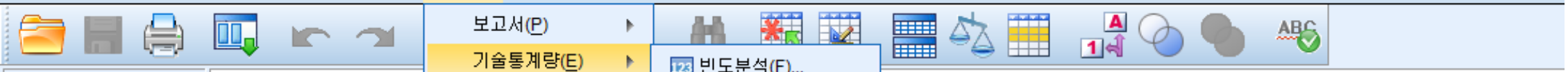
교차분석(카이제곱 검정)

SPSS에서 교차분석을 위해서는 CROSSTABS 명령어를 사용한다.

CROSSTABS는 한정된 수의 고유 값을 갖는 둘 이상의 변수의 결합 분포를 보여주는 빈도 테이블을 생성한다. 하나의 변수의 빈도 분포는 하나 이상의 변수의 값에 따라 세분화된다. 둘 이상의 변수에 대한 고유한 값 조합은 셀을 정의한다.

bank.sav [데이터집합1] - IBM SPSS Statistics Data Editor

파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 창(W) 도움말(H)



분석(A) 메뉴:
- 보고서(P)
- 기술통계량(E)
- 표
- 평균 비교(M)
- 일반선형모형(G)
- 일반화 선형 모형(Z)
- 혼합 모형(X)
- 상관분석(C)
- 회귀분석(R)
- 로그선형분석(O)
- 신경망(W)
- 분류분석(Y)
- 차원 감소(D)
- 척도(A)
- 비모수 검정(N)
- 예측(T)
- 생존확률(S)
- 다중응답(U)
- 결측값 분석(V)...
- 다중 대입(T)
- 복합 표본(L)
- 시뮬레이션...
- 품질 관리(Q)
- ROC 곡선(V)...

기술통계량(E) 하위 메뉴:
- 빈도분석(F)...
- 기술통계(D)...
- 데이터 탐색(E)...
- 교차분석(C)...
- 비율(R)...
- P-P 도표(P)...
- Q-Q 도표(Q)...

	age	job	nce	housing	loan	contact	day	month	du
1	30	unemployed	1787	no	no	cellular	19	oct	
2	33	services	4789	yes	yes	cellular	11	may	
3	35	management	1350	yes	no	cellular	16	apr	
4	30	management	1476	yes	yes	unknown	3	jun	
5	59	blue-collar	0	yes	no	unknown	5	may	
6	35	management	747	no	no	cellular	23	feb	
7	36	self-employed	307	yes	no	cellular	14	may	
8	39	technician	147	yes	no	cellular	6	may	
9	41	entrepreneur	221	yes	no	unknown	14	may	
10	43	services	-88	yes	yes	cellular	17	apr	
11	39	services	9374	yes	no	unknown	20	may	
12	43	admin.	264	yes	no	cellular	17	apr	
13	36	technician	1109	no	no	cellular	13	aug	
14	20	student	502	no	no	cellular	30	apr	
15	31	blue-collar	360	yes	yes	cellular	29	jan	
16	40	management	194	no	yes	cellular	29	aug	
17	56	technician	4073	no	no	cellular	27	aug	
18	37	admin.	2317	yes	no	cellular	20	apr	
19	25	blue-collar	-221	yes	no	unknown	23	may	
20	31	services	132	no	no	cellular	7	jul	
21	38	management	0	yes	no	cellular	18	nov	
22	42	management	16	no	no	cellular	19	nov	
23	44	services	106	no	no	unknown	12	jun	

숙제

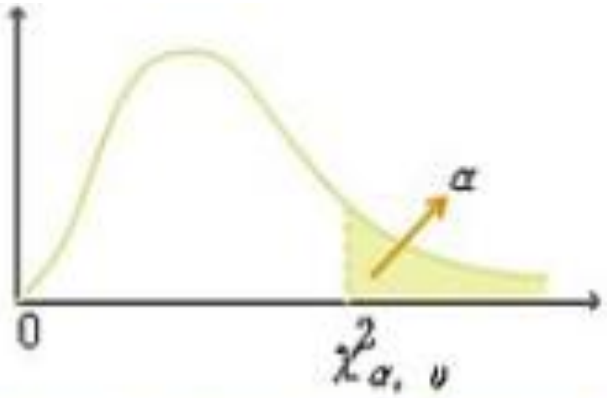
- 숙제1
  - 엑셀로 내보내기
  - 엑셀에서 계산해서 카이제곱 값과 일치하는지 확인하기
    (SPSS값과 일치하는 것 확인하기)

- 숙제2
  - Y와 다른 명목형 변수들간의 교차 분석을 SPSS로 실시하고 결과 해석하기(*.xls 파일로 저장해서, 해석달기)

카이제곱 분포표

χ^2

<부표-3> X^2 분포표



ν	α=.995	α=.99	α=.975	α=.95	α=.05	α=.025	α=.01	α=.005	ν
1	.3333930	.000157	.000982	.00393	3.841	5.024	6.635	7.879	1
2	.0100	.0201	.0506	.103	5.991	7.378	9.210	10.597	2
3	.0717	.115	.216	.352	7.815	9.348	11.345	12.838	3
4	.207	.297	.484	.711	9.488	11.143	13.277	14.860	4
5	.412	.554	.831	1.145	11.070	12.832	15.086	16.750	5
6	.676	.872	1.237	1.635	13.592	14.449	16.812	18.548	6
7	.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11

결과해석하기

- 상관분석
  - Age와 balance의 상관계수는 ###, 유의확률은 ###이므로 유의수준 0.05에서 상관성이 존재하다. 다만, 상관계수값이 작으므로 아주 작은 (+, -)의 상관관계가 존재한다고 할 수 있다.

- 교차분석(CrossTab)
  - 카이제곱(X^2) 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 [Y 집단]별로 [결혼상태]는 서로 다르다(즉, 서로 연관성이 있다)고 할 수 있다.

- 평균차이분석(Means)
  - F통계량 값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 [Y 집단]별로 [duration]은 통계적으로 유의한 차이가 있다.

- 회귀분석
  - 분산분석표에 따라 F 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 회귀모형이 통계적으로 유의하다고 할 수 있다.
  - t 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 회귀계수는 0이 아니다. 즉, [나이]는 [balance]에 영향을 준다(나이에 따라 balance는 달라진다)고 할 수 있다.