10 연관성 분석

01 연관성 분석의 기본

- `연관성 분석`: 조사 대상에서 수집한 자료의 척도를 기준으로 변수들 간에 어느 정도 밀접한 관계가 있는지를 판단하기 위한 분석 방법.
    - 자료의 척도를 기준으로 연관성을 파악=> 척도에 따라 분석 방법도 달라짐.
- 척도에 따른 연관성 분석 구분표

구분	척도	분석 방법	기타 변수 개입 여부
상관분석	등간척도, 비율척도	편상관분석	o
		피어슨 상관분석	x
	서열척도	스치어만 서열 상관분석	-
교차분석	명목척도	교차분석	-

- 연속형 변수들 사이의 연관성 분석법
    - 기타변수 개입O: 편상관분석
    - 기타변수 개입X: 피어슨 상관분석
- 범주형 변수
    - 서열로 구성된 변수: 스피어만 서열 상관분석
    - 명목척도로 구성: 교차분석을 통한 검정

02 상관분석

상관분석의 개념

- 상관분석: 조사 목적에 맞게 구성된 변수들 간의 연관성을 분석하는 방법
- 상관관계 2개의 변수를 기준으로 양, 음의 방량으로 일정한 규칙이 나타나는 선형관계의 형태와 연관 정도를 수치로 나타냄
- 상관계수: 연관성을 나타내는 수치
    - 반드시 두 변수 간의 1:1관계가 수치로 나타남.

산포도

- 산포도: 2개의 변수를 x와 y의 그래프로 나타내 분포 정도를 확인 한것
- 산포도를 나타내는 지표: 분산, 표준편차, 범위, 사분위수, 백분위수
    - 이러한 수치들은 표본내 흩어진 정도-> 변수들간 연관성 알수 X
    - 따라서 연관성을 수치적으로 알기 위해 공분산솨 상관계수에 대해 알아야한다.
- 공분산: 두 가지 확률변수에 대한 흩어짐 정도가 동일한 방향인 양의 방향인지, 반대방향인 음의 방향인지를 나타내는 수치
    - X와 Y값의 단위가 다른경우 공분산 값을 둘의 표준편차의 곱으로 나누어 단위의 효과를 상쇄하여 표준화한다. 이 값이 상관계수이다.

$$\mathrm{X}\text{와 }\mathrm{Y}\text{라는 }2\text{개의 변수 존재, 변수가 평균}(\bar{X},\bar{Y})\text{로부터 떨어져 있는 정도는}$$
$$\sum(X-\bar{X})(Y-\bar{Y})\text{를 표본 개수로 나누어 계산하고, 이를 공분산이라 한다.}$$

03 공분산과 상관계수

공분산

- 공분산: 두 확률변수의 흩어진 정도가 양의 방향인지 반대인 음의 방향인지를 나타내는 수치

$$\text{공분산은 }\mathrm{x}\text{와 }\mathrm{y}\text{의 변화정도 표현, 범위}: -\infty \le Cov \le \infty$$
$$\text{따라서 정도의 차이만 파악, 강도는 확인 불가능 하다.}$$

- 확률변수 X,Y에 대해 흩어짐의 정도가 산포도이며, 이를 분산으로 표시가능.
    - 이 두 확률변수의 공통점이 공분산이며, 이에 대한 분석을 공분산 분석이라 함.
    - 아래식에서 X에 대한 평균편차와 Y에 대한 평균편타의 곱을 모두 합하여 총 관측치의 수로 나눈 값이 공분산이다.

$$\text{공분산은 }Cov(X,Y)\text{로 표시하며 다음과 같이 표현된다.}$$
$$\begin{aligned} Cov(X,Y) &= \frac{\sum_{i=1}^{N}(X-\bar{X})(Y-\bar{Y})}{N} \\ (\text{공분산}) &= \frac{[((\text{개별}X\text{측정치})-(X\text{의평균}))*((\text{개별}Y\text{측정치})-(Y\text{의평균}))]}{(\text{조합을 이루는 개수})} \\ &= \frac{[(X\text{의평균편차})*(Y\text{의평균편차})]\text{의총합}}{(\text{조합을 이루는 개수})} \end{aligned}$$

공분산의 개념

- X와 Y가 같은 방향: 똑같이 +값이거나 -값으로 서로 대응하는 경우 공분산은 커짐.
- X와 Y가 다른 방향: 서로 +값과 -값으로 대응하는 경우 공분산은 작아진다.
- X와 Y가 일정한 규칙 없이 대응: 공분산은 0에 가까워진다.

상관계수

- 상관계수: 공분산을 표준화한 값으로, 강도를 알기 위해서.
    - 상관 계수의 범위는 -1=<Corr=<1로 하안과 상한이 고정 따라서 양과 음의 정도에 대한 파악과 함께 연관성의 강도까지 확인할 수 있다.
    - 공분산만 분석했을 때 관계를 정확히 파악하기 어렵기 때문에 이러한 문제의 극복을 위해 표준화하며 표준화된 공분산 계수를 상관계수라 한다.

$$\begin{aligned} Cov(X,Y) &= \frac{\sum_{i=1}^{N}(X-\bar{X})(Y-\bar{Y})}{N} \\ Corr(x,y) &= \frac{Cov(x,y)}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{n}*\frac{\sum(y_i-\bar{y})^2}{n}}} \\ &= \frac{Cov(x,y)}{S.D(x)*S.D(y)}(\mathrm{S.D(x)}:\mathrm{x}\text{의 표준편차}, \mathrm{S.D(y)}:\mathrm{y}\text{의 표준편차}) \\ (\text{상관계수}) &= \frac{(\text{공분산})}{x\text{의표준편차}*y\text{의표준편차}}\text{공분산을 }\mathrm{x}\text{의 표준편차와 }\mathrm{y}\text{의 표준편차를 곱한 값으로 나눈 값이 상관계수이다.} \end{aligned}$$

상관계수의 개념

- x와 y가 서로 양의 상관관계: 0<p(x,y)=<1의 상관계수 값을 가짐
- x와 y가 서로 음의 상관관계: -1=<p(x,y)<0의 값을 가짐
- x와 y가 일정한 규칙 없이 양,음값이 동시에 대응하면 상관계수=0

상관계수의 가설검정

- 사실관계를 나타내기 위해 표본을 통해 표본상관계수(r)로 모상관계수(p)를 추정하기 위해 추가분석을 하는 과정을 상관계수의 가설 검정이라 한다.

가설 수립

- 상관계수는 0으로 갈수록 상관관계가 0,-1이나 1로 갈수록 연관성 높다.

$$H_0 : p = 0 => \text{연관성이없다.}$$
$$H_1 : p \neq 0 => \text{연관성이있다.}$$

검정통계량

- 상관계수의 검정통계량은 t분포를 이용 (분산을 알지 못하는 상황에서 상관계수의 평균에 대한 표본분포를 확인하기 때문)

$$t_{(n-2)} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = r\sqrt{\frac{n-2}{1-r^2}}$$

  검정통계량 t값이 임계치보다 작으면 귀무가설을 채택, 임계치보다 크면 귀무가설을 기각하고 대립가설을 채택.

예제

$$H_0 : p = 0 => \text{연관성이없다.}$$
$$H_1 : p \neq 0 => \text{연관성이있다.}$$

n=28, 임계치= 2.0484=> t값이 임계치보다 작으면 귀무가설을 채택, 크면 대립가설 채택

$$t_{(28)} = 0.888\sqrt{\frac{28}{1-0.888^2}} = 10.218$$

임계치보다 검정통계량이 크므로 귀무가설을 기각하고 대립가설을 채택, 연관성이 있다.

04 교차분석

교차분석과 카이제곱 검정

- 교차분석: 범주형 척도로 구성된 자료들 간의 연관관계를 확인하기 위해 교차표를 만들어 관계를 분석하는 방법
  - 변수들의 빈도를 확인, 빈도를 이용하여 상호 연관성을 판단
  - 검정통계량으로 카이제곱 검정을 이용.

교차표

- 교차표: 2개의 조사 요인에 대한 자료값을 각각 행과 열로 배열하여 교차되는 항목에 대한 빈도를 나타낸 표

관측빈도와 기대빈도

- 관측빈도: 실제로 수집된 데이터의 빈도 O_ij로 표기
- 기대빈도: 전체 빈도 n에 대하여 행과 열의 합을 기준으로 각 교차되는 셀에 몇번의 빈도가 확인될 수 있을지를 예상하는 기대값.

$$E_{ij} = \frac{n_i * n_j}{n} (E_{ij} = \frac{n_i}{n}, n_j = \text{열의 빈도})$$

카이제곱 통계량

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- 관측빈도와 기대빈도 편차의 제곱을 기대빈도로 나눈값의 합

카이제곱분포의 자유도

$$\text{카이제곱 검정은 범주형 변수를 대상으로 연관성을 판단} \\ d.f(\text{자유도}) = k - 1(k = \text{범주형 변수의 수})$$

- 자유도와 유의수준을 기준으로 작으면 채택, 크면 기각한다.

적합도 검정

- 양자택일의 기대빈도는 50:50으로 예상이 가능하다
- 기대빈도와 관측빈도의 차이가 적으면 적을수록 적합한 기대.
- 카이제곱 분포를 이용하여 차이가 있는지 없는지를 검정할 수 있다.
- ex)

$$H_0: \text{바다에 대한 선호도}=\text{산에 대한 선호도} \\ H_1: \text{바다에 대한 선호도} \neq \text{산에 대한 선호도} \\ \chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(68-50)^2}{50} + \frac{(32-50)^2}{50} = 12.96$$

임계치는 3.8이므로 검정값이 임계치보다 큰 기각역에 속하여 귀무가설을 기각, 대립가설을 채택한다.

독립성 검정

- 독립성 검정: 여러가지 범주를 대상으로 각 범주가 독립적인지를 판단하는 검정법.
- 독립성 검정에서의 자유도는 다음과 같다.

$$d.f = (R-1)(C-1)\ (\mathrm{R}:\text{행의 수}, \mathrm{C}:\text{열의 수})$$

- ex)

$$H_0: \text{지역과 구매 의사는 독립적이다.} \\ H_1: \text{지역과 구매 의사는 독립적이지 않다.} \\ d.f = (2-1)(2-1) = 1 \\ \chi^2 = 14.407, \alpha = 6.63$$

카이제곱값이 임계값보다 크므로 기각역에 속하여 귀무가설을 기각하고 대립가설을 채택, 지역과 구매 의사는 독립적이지 않다.

연습문제

4. 공분산은 -11.18로 음의 상관관계를 나타낸다.
5. 상관계수는 -0.867로 음의 상관관계이다. n=23, t=8.344이므로 임계치 2.0686보다 크므로 연관성이 있다고 볼수 있다.
6. $$H_0: \text{폴더블 선호도}=\text{롤러블 선호도} \\ H_1: \text{폴더블 선호도} \neq \text{롤러블 선호도} \\ E_{ij} = 7.5, d.f = 1, \chi^2 = 0.6$$

   임계치는 3.8415로 카이제곱값보다 크다. 따라서 귀무가설을 채택하므로 둘의 선호도는 같다.
7. cell을 보자