1. 상관관계분석이란?

1) 개념

두 변수의 관계를 설명할 때 가장 기본적인 설명 방법이 관련성 정도를 설명하는 것인데, 두 변수가 명목변수인 경우에는 교차분석을, 등간 이상의 연속 확률변수인 경우(주로 비율변수)에는 상관관계분석을 통해 통계적 관련 정도의 유의미성을 분석한다.

예를들어, 우리나라 자치시의 인구증가율과 합계출산율의 관계를 분석한다고 가정하자. 두 변수는 모두 비율변수이므로 상관관계분석을 통하여 결론을 도출하게 된다. 결론은 세가지고 나뉜다.
1. 양의 관계
2. 음의 관계
3. 관련성 없거나 매우 약한 경우

2) 기본가정

상관관계분석을 하기위한 기본가정은 다음과 같다.

1. 두 변수는 등간변수나 비율변수여야 한다.
2. 두 변수 간의 관계는 선형 함수 관계이어야 한다.
3. 두 변수의 분산이 동일하여야 한다.
4. 두 변수의 분포는 정규분포이여야 한다.
5. 표본을 대상으로 상관관계분석을 시행한다면 포본의 크기는 충분히 커야한다.

2. 상관계수의 이해

1) 상관계수의 계산

이 장에서 다룰 상관계수는 피어슨의 상관계수이다. 피어슨 상관계수는 선형 관계인 두 확률 변수의 상관성 정도를 추정할 수 있는 계수이다. 수학적으로 상관계수는 공분산을 두 변수의 표준편차로 나누어 도출한다. 먼저 공분산은 다음과 같이 구한다.

$$\text{두 변수의 공분산} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

두 변수가 같은 방향으로 움직이는 경향이 커질수록 공분산도 커진다. 하지만 공분산은 두 변수의 단위에 영향을 받는다. 다시 말해 단위가 커지면 공분산 값도 커진다. 따라서 이 값을 두 변수의 표준편차로 표준화할 필요가 있다. 두 변수를 a와 b라고 했을 때 상관계수는 다음과 같다.

$$a\text{와 } b\text{의 상관계수} = \frac{cov(a,b)}{sd(a) \times sd(b)}$$

이 값을 상관계수라고 하고 -1에서 1사이의 값이다.

2) 상관계수의 의미

상관계수는 일반적으로 r로 표시하고 [-1,1]사이의 값이다. 상관계수가 음수이면 두 변수의 관계가 부정관계이고, 양수이면 긍정관계이다.
r이 1이라면 두 변수는 같은 변수이고 -1이면 두 변수는 역의 관계로 같은 변수이다.

3) 가설의 설정

상관관계분석에서 귀무가설은 “비율변수인 x와 비율변수인 y는 상관성이 없다.”이고 대립가설은 “x와 y가 상관성이 있다.”이다.

상관계수이 값이 0이면 상관성이 없는 상태를 의미하므로 귀무가설은 $H_0 : r_{x,y} = 0$으로, 대립가설은 $H_0 : r_{x,y} \neq 0$로 정의된다.

3. R 활용 상관관계분석

```
> setwd("C:/Users/USER/Desktop/교육원초급/data/xlsx 파일")
> library(openxlsx)
> mydata5 = read.xlsx("cor.xlsx",1) # 파을을 불러와 mydata5로 저장
> attach(mydata5) # 불러온 데이터를 R과 연동
> View(mydata5) # 불러온 데이터를 R Studio에서 확인
```

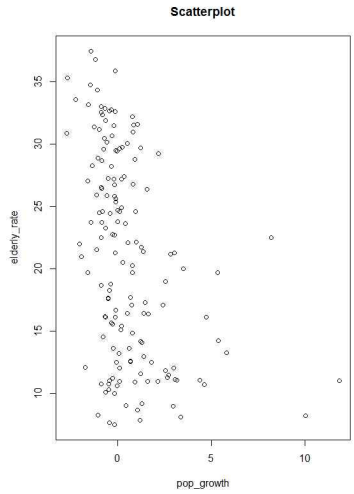
군별	pop_growth	birth_rate	elderly_rate	finance	cultural_center	dummy
경기 수원시	1.06	1.292	8.71	0.729	2.79	1
경기 성남시	-0.26	1.159	11.26	0.768	2.16	1
경기 의정부시	0.69	1.104	12.63	0.606	1.38	1
경기 안양시	-0.47	1.177	10.32	0.721	2.34	1
경기 부천시	-0.64	1.072	10.13	0.662	2.94	1
경기 광명시	-0.88	1.235	10.77	0.691	2.32	1
경기 평택시	2.67	1.469	11.28	0.636	2.61	1
경기 동두천시	0.52	1.292	16.46	0.575	4.08	1
경기 안산시	-1.05	1.219	8.28	0.666	3.44	1
경기 고양시	2.14	1.161	10.97	0.659	3.02	1

```
> names(mydata5) # 불러온 데이터에 포함된 변수의 이름을 확인
[1] "군별"           "pop_growth"     "birth_rate"     "elderly_rate"
"finance"        "cultural_center" "dummy"
```

이 데이터에서 pop_growth는 인구증가율, birth_rate는 합계출산율, elderly_rate는 65세 이상 노령인구의 비율을 의미한다. 그리고 finance는 지방자치단체의 재정자주도를, cultural_center는 문화기반 시설의 수, dummy는 해당 표본이 시인지 군인지를 구분해주는 변수이다.

인구증가율과 노령인구 비율의 산점도와 상관계수를 도출한다.

```
> plot(pop_growth, elderly_rate,main="Scatterplot") # 두 변수의 그래프 그리기
```




```
> cor(pop_growth, elderly_rate,method="pearson") # 두 변수의 상관계수 구하
[1] -0.4069218
```

인구증가율과 노령인구 비율의 상관계수가 -0.41이므로 두 변수는 약한 음의 상관관계에 있다. 또한 산점도를 통해 두 변수가 음의 관계에 있는 것을 확인할 수 있다.

다음으로 다변량 상관관계분석을 시행하기 위해 dummy라는 명목변수를 제외시켰다.

```
> x = cbind(pop_growth,birth_rate,elderly_rate,finance,cultural_center)
> cov(x) # x의 공분산 행렬 구하기
                pop_growth   birth_rate elderly_rate     finance cultural_center
pop_growth      4.20123170  0.080548416   -6.8398841  0.025198983     -3.83992032
birth_rate      0.08054842  0.059188476    0.1131702 -0.001293588      0.19522097
elderly_rate   -6.83988413  0.113170155   67.2510370 -0.190349688     36.95927513
finance         0.02519898 -0.001293588   -0.1903497  0.002837680     -0.02895507
cultural_center -3.83992032  0.195220975   36.9592751 -0.028955067     81.62113529
> cor(x) # x의 상관관계 행렬 구하기
                pop_growth   birth_rate elderly_rate     finance cultural_center
pop_growth      1.0000000  0.16152885   -0.40692178  0.23078789     -0.20736363
birth_rate      0.1615289  1.00000000    0.05672361 -0.09981501      0.08881914
elderly_rate   -0.4069218  0.05672361    1.00000000 -0.43573354      0.49885306
finance         0.2307879 -0.09981501   -0.43573354  1.00000000     -0.06016467
cultural_center -0.2073636  0.08881914    0.49885306 -0.06016467      1.00000000
```

다변량 상관관계분석 결과 노령인구비율과 문화기반시설이 중간 정도의 양의 상관성을 보이고, 인구증가율과 노령인구비율은 약한 음의 상관관계를 보였다. 또한 다른 변수들에서는 상관관계가 뚜렷하지 않았다.