제 11 장. 기초통계분석

1

상관관계 분석

1. 상관관계분석의 의의
2. 기본모형
3. 계산절차
4. 상관관계분석의 사례

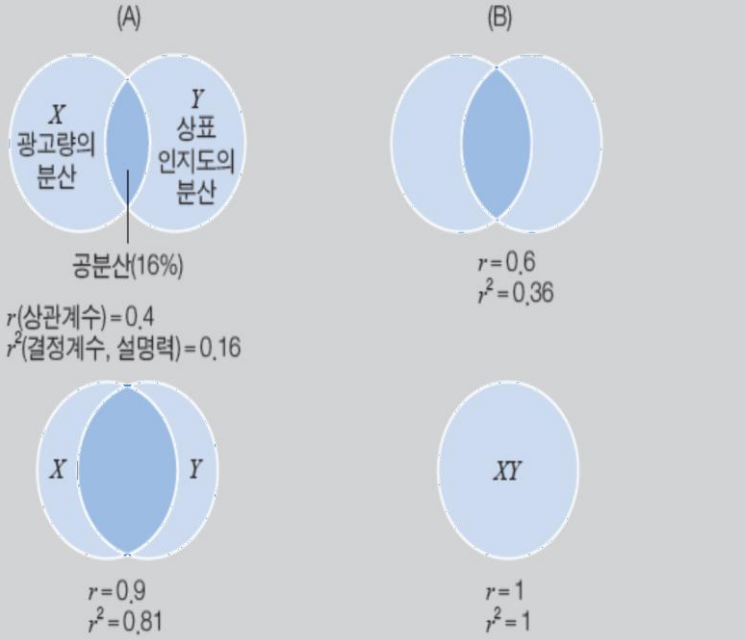
상관관계 분석의 의의(correlation analysis)

(1) 사용목적 : 연구하고자 하는 변수들 간의 관련성을 분석하기 위해서 사용

(2) 기본원리 : 변수들 간 관련성의 정도는 특정변수의 분산 중에서 다른 변수와 같이 변화하는 분산(공분산)이 어느 정도인가에 따라 좌우됨

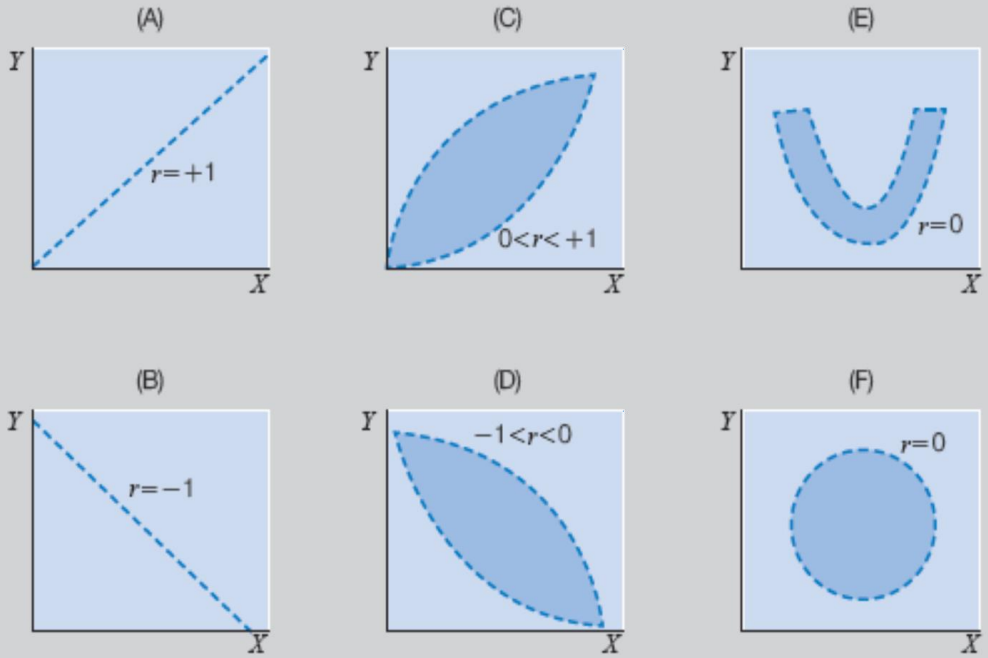
【그림 11-1】

공분산과 상관계수의 관계



【그림 11-2】

상관관계의 종류



상관관계 분석의 의의(correlation analysis)

(3) 상관관계의 종류

✓ 단순상관계수(simple correlation coefficient)

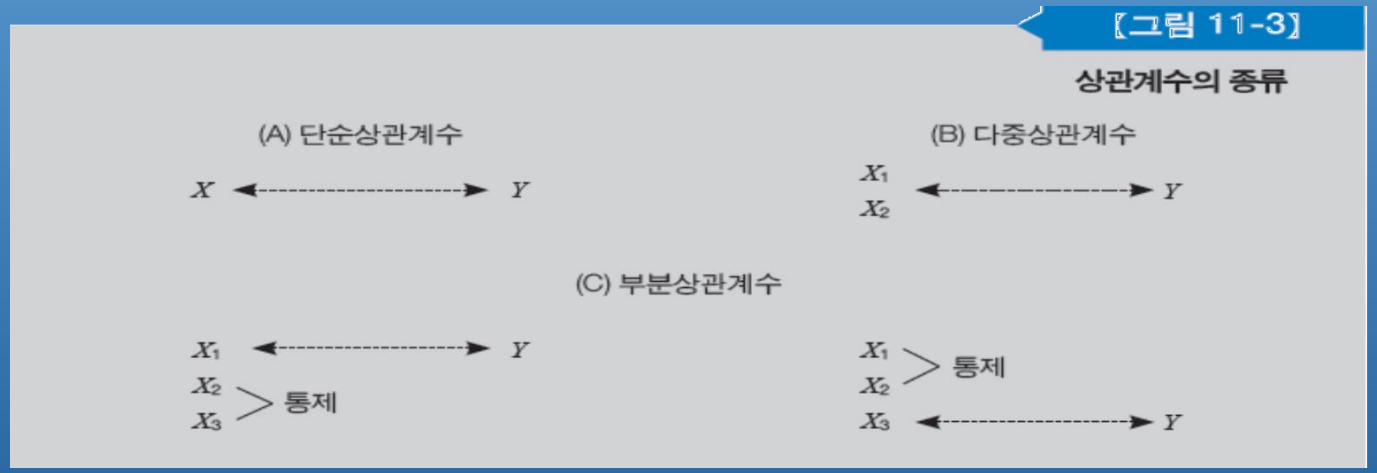
- 두 변수간의 상관관계를 나타내는 계수

✓ 다중상관계수(multiple correlation coefficient)

- 두 변수 이상의 변수 간 상관관계를 나타내는 계수

✓ 부분상관계수(partial correlation coefficient)

- 다른 변수들의 상관관계를 통제하고(다른 변수들과 같이 변화하는 부분은 제외하고) 순수하게 두 변수간의 상관관계를 나타내는 계수



기 본 모 형

✓ 공분산(covariance)

- 확률변수 X의 증감에 따른 확률변수 Y의 증감에 대한 척도

$$\mathrm{E}[(\mathrm{X}-\bar{X})(\mathrm{Y}-\bar{Y})] \;=\; \mathrm{Cov}(\mathrm{X},\mathrm{Y})$$

✓ 모집단상관계수(population carrelation coefficient)

- 두 변수 사이의 선형관계의 강도를 나타내 줄 수 있도록 공분산을 표준화한 것

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

✓ 표본상관계수(sample correlation coefficient)

- 모집단상관계수의 추정치

$$r = \frac{S_{XY}}{S_X S_Y}$$

계 산 절 차

소비자	연령(X)	청량음료 주간소비량(Y)
1	40	84
2	20	120
3	60	24
4	50	36
5	45	72
6	30	84
7	25	96
8	35	48
9	30	60
10	45	36

✓ 예시

- 청량음료 생산업체 성북음료는 신제품을 개발하고 효과적인 마케팅 전략을 세우기 위해 소비자 연령에 따른 청량음료 소비량을 조사하여 왼쪽과 같은 결과를 얻었다.
  성북음료의 마케팅 조사자는
  연령과 청량음료 소비량 사이에 관계가 있는지,
  또 있다면 얼마나 강한 관계를 형성하고 있는지를 파악하고자 한다.

계 산 절 차

소비자	연령(X)	청량음료 주간소비량(Y)
1	40	84
2	20	120
3	60	24
4	50	36
5	45	72
6	30	84
7	25	96
8	35	48
9	30	60
10	45	36

$$r = \frac{S_{XY}}{S_X S_Y}$$

$$= \frac{\sum(X - \bar{X})\ (Y - \bar{Y})}{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2} = -0.833$$

✓ 예시에서 상관관계의 의미

- 소비자 연령과 청량음료 사이 강한 역의 상관관계 존재
- 청량음료의 주 소비층이 청소년임을 확인하고 청소년을 주 대상으로 마케팅 전략 수립 가능

상관관계분석의 사례

	외식빈도	소유한 카드수
1	4	2
2	6	2
3	6	4
4	7	4
5	8	5
6	7	5
7	8	6
8	10	6
9	4	2
10	6	2

✓ 예시

- 오성 카드 회사는 카드 수와 외식빈도 사이에 관계가 있는지, 있다면 어느 정도의 관련성을 지니는지 알아보고자 왼쪽의 자료를 이용해 상관관계분석을 실시하였다.

상관관계분석의 사례

【표 11-1】 상관관계분석 결과

		외식 빈도	카드 수
외식 빈도	Pearson Correlation	1	0.867**
	Sig. (2-tailed)		0.001
	N	10	10
카드 수	Pearson Correlation	0.867**	1
	Sig. (2-tailed)	0.001	
	N	10	10

**Correlation is significant at the 0.01 level(2-tailed).

2

t-test

1. t-test의 의의
2. 독립표본 t – test
3. 동일표본 t – test

t–test 의 의의

✓ Z-test와 t-test의 사용기준

- 모집단의 분산을 알고 있는 경우 Z-test 사용
- 모집단의 분산을 알 수 없는 경우 t-test 사용
- 모집단의 분산을 알 수 없는 경우라 할지라도 표본의 크기가 30개를 초과하면 중심극한정리에 따라 정규분포를 가정할 수 있으므로 Z-test 사용 가능

✓ t-test의 의의

- 독립된 두 개의 표본평균간의 차이를 검증하는 분석방법
  즉, 두 집단간의 평균이 통계적으로 유의한 차이를 보이는지 여부를 검증할 때 사용
- 대응표본 t-test(Paired t-test)
  - 동일한 표본에서 두 개의 변수의 평균값을 비교할 때 사용하는 방법
- 비교집단이 세 개 이상이 되면 t-test의 사용이 불가능하며 분산분석(ANOVA)를 사용해야 한다.

t–test 의 의의

【표 11-2】 두 집단 비교분석방법

모집단 표준편차	표본크기(n)	
	$n \leq 30$	$n > 30$
알 수 있을 때	Z-테스트 $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Z-테스트 $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
알 수 없을 때	t-테스트 $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $S^2 = \frac{\sum_{i=1}^{n_1}(x_1 - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$	Z-테스트 $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $S^2 = \frac{\sum_{i=1}^{n_1}(x_1 - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$

독립표본 t-test

(1) 독립된 두 집단의 t-test의 계산

✓ t-test의 기본 원리

- 각 표본의 분산과 두 표본을 합산한 전체집단의 분산을 이용하여 두 집단간 평균이 통계적으로 유의한 차이가 있는가를 검증하는 것

공식-1 (두 모집단의 분산이 상이)	공식-2 (두 모집단의 분산이 동일)
$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}}$	$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{{S_p}^2}{n_1} + \frac{{S_p}^2}{n_2}}}$
${S_i}^2 = \frac{\sum(X_i - \bar{X}_i)^2}{n_i - 1}$	${S_p}^2 = \frac{(n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2}{n_1 + n_2 - 1}$

$n_1, n_2 =$ 집단 1, 2의 표본수

$S_1, S_2 =$ 집단1, 2의 표본분산

독립표본 t-test

(1) 독립된 두 집단의 t-test의 계산

(단위 : 만원)

번호	남성	번호	여성
1	75	1	51
2	87	2	70
3	83	3	37
4	45	4	62
5	95	5	90
6	89	6	72
7	74	7	45
8	110	8	78
9	75	9	45
10	84	10	76

✓ 예시(두 집단의 분산이 동일하다는 가정)

- K회사는 신제품을 출시하면서 대대적인 광고캠페인을 벌였다. 어느 정도 시간이 경과한 후 성별에 따라 차별적인 광고전략을 세우기 위해 성별에 따라 매출액에 차이가 있는지 여부를 알아보고자 한다.

독립표본 t-test

(1) 독립된 두 집단의 t-test의 계산

(단위 : 만원)

번호	남성	번호	여성
1	75	1	51
2	87	2	70
3	83	3	37
4	45	4	62
5	95	5	90
6	89	6	72
7	74	7	45
8	110	8	78
9	75	9	45
10	84	10	76

✓ 검정가설

- H_0 : 광고캠페인 후 남성과 여성의 매출액은 동일하다.

  H_1 : 광고 캠페인 후 남성과 여성의 매출액은 동일하지 않다.

✓ 공식-2 이용

$S_1^2 = 284.6779$

$S_2^2 = 302.2660$

$S_p^2 = 293.4720$

$t = 2.49$

독립표본 t-test

(2) SPSSWIN을 이용한 독립된 두 집단의 t-test

【표 11-3】 독립된 두 집단의 t-테스트 분석 결과

(단위: 만 원)

	성별	N	Mean	Std. Deviation	Std. Error Mean
매출액	남자	10	81.7000	16.87240	5.33552
	여자	10	62.6000	17.38582	5.49788

【표 11-4】 독립표본 테스트

		Levene's Test for Equality of Variance		t-test for Equality of Means						
		F	Sig.	t	df	sig. (2-tailed)	Mean (Difference)	Std. Error (Difference)	95% Confidence interval of the Difference	
									Lower	Upper
매출액	등분산 가정 값	0.458	0.507	2.493	18	0.023	19.10000	7.66123	3.00435	35.19565
	등분산이 가정되지 않은 값			2.493	17.984	0.023	19.10000	7.66123	3.00332	35.19668

동일표본 t-test

(1) 동일표본 t-test의 계산

✓ 두 집단의 평균을 비교하는 것보다는 동일한 집단에서 두 변수간 평균을 구하는 데 관심

	실시 전 점수		실시 후 점수
1	68	1	70
2	62	2	62
3	50	3	54
4	75	4	82
5	76	5	75
6	57	6	64
7	60	7	58
8	53	8	57
9	74	9	80
10	60	10	63

✓ 예시(두 집단의 분산 동일 가정)

- 한 입시학원에서는 새로운 교육프로그램을 개발했다. 학원 관계자는 새로운 교육프로그램이 학생들의 학업성적을 향상시키는 데 성과를 거둘 수 있을 것인지 알아보기 위해 이 교육프로그램을 실시하기 전과 실시한 후의 학생들의 학업성적을 비교해 보았다.

✓ 동일표본의 평균차이 검정에 사용되는 공식

$$t = \frac{\bar{d} - d_0}{S_d/\sqrt{n}}$$

$\bar{d}$: 각 표본요소 값들의 차이의 평균값
d_0 : 귀무가설로 설정된 차이의 평균값
S_d : 표본요소들의 차이값들의 평균값

동일표본 t-test

(1) 동일표본 t-test의 계산

	실수 후 점수와 실시 전 점수의 차이
	2
	0
	4
	7
	-1
	7
	-2
	4
	6
	3
평균	3
표준편차	3.231786572

✓ 검정가설

- H_0 : 새로운 교육프로그램의 실시 전과 후의 학생들의 점수는 동일하다.

H_1 : 새로운 교육프로그램의 실시 전과 후의 학생들의 점수는 동일하지 않다.

$$t = \frac{\bar{d} - d_0}{S_d/\sqrt{n}} = \frac{3-0}{3.23/\sqrt{10}} = 2.94$$

독립표본 t-test

(2) SPSSWIN을 이용한 동일표본 t-test

【표 11-5】 동일표본 t-테스트 분석 결과

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	실시 전	63.5000	10	9.31248	2.94486
	실시 후	66.5000	10	9.80079	3.09928

Paired Samples Test

	Paired Differences					t	df	sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 실시 전-실시 후	−3.00000	3.23179	1.02198	−5.31188	−0.68812	−2.935	9	0.017

THANK YOU