SPSS
온라인 강좌


Thanks for visiting
mySTATISTICS.net
We are currently offering
the best service through
statistical analysis to those
We will offer you
a satisfactory service..

1. 척도 (Scale)

척도는 질적인 대상을 양적인 대상으로 전환해주는 도구로써, 관찰 대상을 측정하기 위하여 대상의 속성을 일련의 기호나 숫자로 나타내는 것을 의미한다.

(1) 명목척도 (Nominal scale) : 명목척도는 숫자 자체로써 크기를 나타내는 적인 의미는 없다. 그래서 관찰 대상을 구별하거나 확인하기 위하여 주로 사용된다.

예) 귀하의 성별은? (1) 남자 (2) 여자

→ 위의 경우, 남자, 여자간 숫자 1,2 에 대한 의미는 없다. 숫자 2가 1 보다 2배가 크다거나 하지 않다. 단지 남자는 1, 여자는 2로 표기한 것 일 뿐이다.

(2) 서열척도 (Ordinal scale): 관찰 대상의 순서적인 특성만을 나타낸 것으로, 관찰대상의 비교 우위를 결정할 뿐, 순서간 차이는 의미가 없다.

예) 가장 좋아하는 음식을 좋아하는 순서대로 세가지만 적으세요.
1.( ), 2. ( ), 3. ( )

→ 음식을 좋아하는 순서가 중요한 것이지, 순서간 차이는 별 의미가 없다. 마치 마라톤에서 1위, 2위, 3위 결정 하는데 있어서 1,2,3 위간 시간차이가 중요한 것이 아니라, 순서가 중요한 것과 비슷한 이치이다.

(3) 등간척도 (Interval scale): 관찰대상의 속성에 대한 차이를 균일한 간격으로 나누어서 측정하는 척도이다. 그래서 척도간 차이는 같다고 할 수 있다. 리커트 척도 (Likert scale) 가 이에 해당한다.

예) 다음 신상품에 대하여 귀하께서는 어느정도 만족 하고 계십니까?
(1) 매우 불 만족한다 (2) 불만족 한다 (3) 그저그렇다 (4) 만족한다 (5) 매우 만족한다

→ 만약 소비자 만족도 조사에 5점 척도가 설문지 에서 사용되었다면, 1-2 사이의 간격은, 4-5 사이의 간격과 같아야 한다.

(4) 비율척도 (Ratio scale): 등간척도와 같은 개념이지만, 절대적인 0 을 가지고 있는 척도이다. 그래서 거리, 무게, 시간 등 측정에 사용된다.

예) 귀하의 나이는? ( ) 혹은 귀하의 수입은? ( )

→ 나이 혹은 수입의 경우 0 이라는 절대값이 존재하게 된다. 하지만 위의 등간척도에서 '소비자 만족'의 경우 절대값 0 이 존재하지 않으므로 등간척도와는 차별이 된다.

2.빈도분석 (Frequency Analysis)

Analyze → Descriptive Statistics → Frequencies

빈도분석은 도수 분포표 혹은 막대그래프를 이용하여 변수의 분포특성을 나타내는 통계 기법이다.

● **도수 분포표**

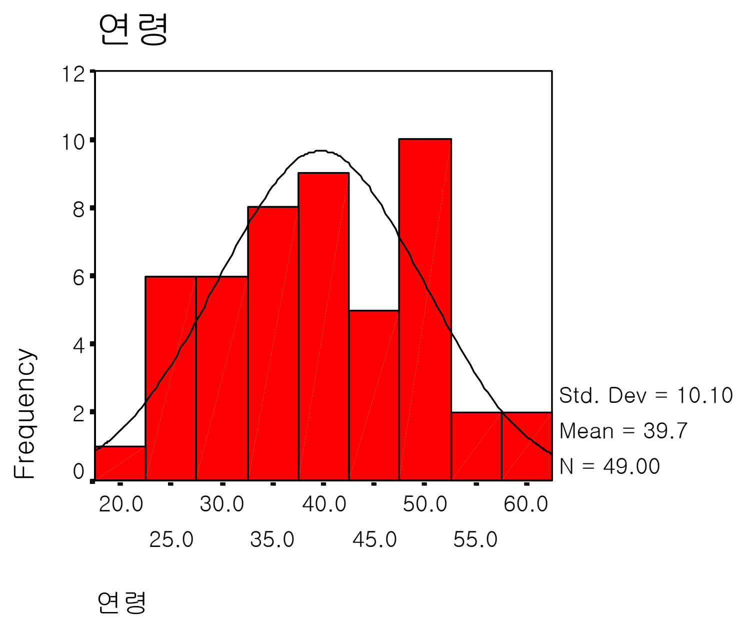
위에 표에서, Valid 는 나이 (20~59세) 를 나타내며, Frequency 는 나이를 나타내는 빈도를 나타낸다.

Percent 는 각 나이별 백분율을 나타내고 있고,
Valid Percent 는 Missing value (무응답, 여기서는 1명) 를 제외한 나머지 (여기서는 49명) 를 유효한 수로 간주하여 그것을 기본으로 나타낸 백분율이고,
Cumulative Percent 는 각 나이당 누적 백분율을 나타낸다.

연령

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20	1	2.0	2.0	2.0
	23	1	2.0	2.0	4.1
	24	4	8.0	8.2	12.2
	27	1	2.0	2.0	14.3
	29	2	4.0	4.1	18.4
	31	3	6.0	6.1	24.5
	32	1	2.0	2.0	26.5
	33	1	2.0	2.0	28.6
	34	1	2.0	2.0	30.6
	35	2	4.0	4.1	34.7
	36	1	2.0	2.0	36.7
	37	3	6.0	6.1	42.9
	38	2	4.0	4.1	46.9
	39	1	2.0	2.0	49.0
	40	2	4.0	4.1	53.1
	41	2	4.0	4.1	57.1
	42	2	4.0	4.1	61.2
	43	1	2.0	2.0	63.3
	44	1	2.0	2.0	65.3
	45	1	2.0	2.0	67.3
	46	2	4.0	4.1	71.4
	48	2	4.0	4.1	75.5
	49	2	4.0	4.1	79.6
	50	1	2.0	2.0	81.6
	51	4	8.0	8.2	89.8
	52	1	2.0	2.0	91.8
	54	1	2.0	2.0	93.9
	56	1	2.0	2.0	95.9
	58	1	2.0	2.0	98.0
	59	1	2.0	2.0	100.0
	Total	49	98.0	100.0	
Missing	System	1	2.0		
Total		50	100.0		

● **히스토그램(histogram)**



위의 그림은 도수 분포표에 나타난 연령을 막대그래프로 나타낸 것으로 전체응답자 (N = 49) 는 49 명이고, 평균 (Mean) 은 39.7 이고, 표준편차는 (Std. Dev) 10.10 세 이다. 여기에서 곡선은 정규곡선이다.

일반적으로 기술통계에서 아래의 Descriptive Analysis에서는 디폴트로 최소값(minimum), 최대값(maximum), 평균(mean), 표준편차(standard deviation), 합계(sum)나타나지만 Frequency analysis의 통계량 option 에서는 해당하는 값을 산출할수 있다. 그리고 여기에서는 Descriptive Analysis 에서는 산출이 불가능한 사분위수 및 백분위수을 산출가능하다.

3. 기술통계분석 (Descriptive Analysis)

Analyze → Descriptive Statistics → Descriptives

기술통계분석은 빈도분석에서 볼 수 있는 통계량 뿐만 아니라, 기타 다른 여러가지 통계량을 쉽게 나타내어 준다.

Descriptive Statistics

	N	Range	inimur	laximu	Mean	Std.	arianc	Skewness		Kurtosis	
	tatistic	tatistic	tatistic	tatistic	tatistic	tatistic	tatistic	tatistic	d. Err	tatistic	d. Err
연령	49	39	20	59	39.69	10.10	2.092	-.055	.340	-.841	.668
Valid N (li	49										

- **기술통계분석 결과표**

위의 표에서는, 빈도분석에서 보여주었던, 전체응답자 (N = 49), 평균 (39.69),표준편차 (10.10) 이외에도, 나이의 최대값 (Maximum), 최소값 (Minimum), 분산 (Variance), 왜도 (Skewness), 첨도 (Kurtosis) 등 다양한 통계량을 제공하고 있다.

- 왜도 (Skewness): 왜도는 분포의 치우침을 나타내는 것으로, 오른쪽꼬리분포는 + 값을, 왼쪽꼬리분포는   값을 보인다.(대칭분포의 왜도=0)

- 첨도 (Kurtosis): 첨도는 평균값을 중심으로 어느정도 밀집 (뾰족하거나 무딘 형태) 되어 있는지를 보여준다. 뾰족한 형태일수록 +, 그 반대는 이다.(표준정규분포의 첨도=0).

또한 여기에서는 표준화된 Z값을 저장할 수 있다.

3.카이스퀘어 (교차분석)

Analyze → Descriptive Statistics → Crosstabs
Option: Statistics→ Chi-square
Cells → Row, Column, Total

카이스퀘어 검정은 두변수가 상호독립적인지 아니면 관련성(association)이 있는지 여부를 알아보는 방법이다.
사용되는 척도 는 모두 명목척도이거나 순서척도이며, 구간이나 비율척도는 변수값의 가지수가 몇 안되거나 코딩변경(recoding)을 통한 그룹화한 변수로 이용가능하다.

교차분석은 독립성 검정(Test of Independent)과 동질성 검정(Test of Homogeneity)로 구분하고 있으나 여기에서는 독립성 검정으로 함께 취급한다.

귀무가설(H0) : 두변수는 서로 독립적이다.
대립가설(H1) : 두 변수는 연관성이 있다. 혹은 한 변수값에 따라 다른 변수값의 패턴이 다르다.

로 설정된다. 즉, 귀무가설이 사실이라면 관측빈도와 기대빈도가 대부분이 비슷하게 예측되는 것을 의미한다

예를 들어서 학력과 담당업무에 대해 서로 연관이 있는지 없는지를 알아보기

위하여 다음과 같은 분석을 실시 하였다. (표는 다음 페이지)

학력 * 담당업무 교차표

			담당업무			전체
			강의	행정관리	프로그램개발	
학력	고졸이하	빈도	0	4	0	4
		기대빈도	.4	3.0	.6	4.0
		학력의 %	.0%	100.0%	.0%	100.0%
		담당업무의 %	.0%	10.8%	.0%	8.0%
		전체 %	.0%	8.0%	.0%	8.0%
	전문대졸	빈도	0	3	0	3
		기대빈도	.3	2.2	.5	3.0
		학력의 %	.0%	100.0%	.0%	100.0%
		담당업무의 %	.0%	8.1%	.0%	6.0%
		전체 %	.0%	6.0%	.0%	6.0%
	4년재 대졸	빈도	0	16	4	20
		기대빈도	2.0	14.8	3.2	20.0
		학력의 %	.0%	80.0%	20.0%	100.0%
		담당업무의 %	.0%	43.2%	50.0%	40.0%
		전체 %	.0%	32.0%	8.0%	40.0%
	대학원졸이상	빈도	5	14	4	23
		기대빈도	2.3	17.0	3.7	23.0
		학력의 %	21.7%	60.9%	17.4%	100.0%
		담당업무의 %	100.0%	37.8%	50.0%	46.0%
		전체 %	10.0%	28.0%	8.0%	46.0%
전체		빈도	5	37	8	50
		기대빈도	5.0	37.0	8.0	50.0
		학력의 %	10.0%	74.0%	16.0%	100.0%
		담당업무의 %	100.0%	100.0%	100.0%	100.0%
		전체 %	10.0%	74.0%	16.0%	100.0%

각 셀당 통계수치는 위에서부터 순서대로 빈도수, 행%, 열%, 전체% 순으로 수치를 나타내고 있다. 예로써, 대학원졸 이상이면서 행정관리인 응답자는 14명으로, 대학원졸업 이상의 학력자 23명중 14명이 되어 60.9% 이고, 행정관리직 37명중 14명이 되어 37.8% 가 된다.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.490[a]	6	.204
Likelihood Ratio	11.458	6	.075
Linear-by-Linear Association	.289	1	.591
N of Valid Cases	50		

a. 10 cells (83.3%) have expected count less than 5. The minimum expected count is .30.

위의표에서 보여주듯이 Pearson Chi-Square 값이 8.490 이고 자유도가 6일 때, Sig (P) = 0.204 <0.05이므로 유의수준 5% 에서 유의하지 않으므로 학력과 담당업무가 상호 독립적이라는 귀무가설이 채택된다.
다시 말해서, 위의 연구에서는 학력과 담당연구간 관련성이 없다고 말할 수 있다. 따라서 위의 예를 보면 측정빈도와 기대빈도가 각 셀마다 유사함을 알 수 있다.

5. 표본 평균의 검정 (T-test)

표본평균의 검증은 일정한 기준을 바탕으로 평균에 대한 가설을 채택할 것인지 아니면 기각을 판단하는 절차이다. 우선 표본평균의 가설검정은 단일 모집단 평균검증 과 두 모집단 평균차이 검증으로 나누어 진다.

(1) 단일모집단 평균검증

Analyze → Compare means → One-Sample T-test

- **Test value 란에 기준치 입력**

단일 모집단 평균검증은 하나의 모집단에서 추출된 표본에 대해서 검정하는 방법으로써 아래와 같은 예를 들 수 있다.

일반적으로 한국 성인 남성의 몸무게가 70 kg 이 되는 것으로 알려져 있다. 이를 바탕으로 서울시내의 성인 남성의 평균 몸무게가 70kg 이라고 할 수 있는지 알아보기 위하여 서울시내에 거주하는 성인 남성 50명을 대상으로 몸무게를 측정하였다. 그렇다면 과연 이 결과로써 서울시내 성인 남성의 평균 몸무게가 70kg 이라고 말할 수 있는가?

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
몸무게	50	71.6600	9.4752	1.3400

위의 표는 샘플 크기, 평균, 표준편차, 표준오차에 대한 값을 보여주고 있다.

One-Sample Test

	Test Value = 70					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
몸무게	1.239	49	.221	1.6600	-1.0328	4.3528

위의 결과로 보아서, Sig. (2-tailed) = 0.221 > 0.05 이므로 귀무가설이 채택된다. 다시 말해서 응답자의 평균 몸무게는 70 이라고 할 수 있다. 또한 귀무가설의 수치인 70kg 과 표본평균간의 차이는 1.660 이 되겠다.
95% 신뢰구간을 보더라도 신뢰구간인 -1.0328 ~ 4.3528 이 0 을 포함 하므로 귀무가설이 채택된 것이다.

(2) 두 모집단의 평균차이검증

Analyze → Compare means → Independent-Samples T test

사용되는 척도 - 집단변수는 명목척도, 순서척도, 검정변수는 구간, 비율척도 기본가정 두 모집단의 분포는 각각 정규분포를 따른다는 가정하에서 분석을 하며 소표본인 경우는 반드시 모집단의 정규성을 검토한 후 정규분포를 만족하지 못하는 경우는 독립표본 t 검정을 이용할 수 없으며 이 경우는 비모수검정 중 Mann-Whitney U test를 통하여 분석 가능하다. 그러나 대표본(일반적으로 n>30)인 경우는 모집단의 분포에 관계없이 적용가능하다.

두 모집단의 평균차이 검증은 말 그대로 하나의 모집단이 아닌 다른 두 모집단의 평균에 대한 차이를 검증하는 방법으로써, 가장 보편적으로 많이 사용되는 기법이기도 하다. 예를 들자면 아래와 같다.

초등학교 영재들을 대상으로 새롭게 두 가지 교육 방법이 개발되었다. 두 가지 방법 중 어느 방법이 더 효과가 있는지 알아보기 위하여 초등학교 영재학생 40 명을 무작위로 20명씩 두 그룹으로 나누어서 각각 A 방법과 B 방법으로 나누어 교육을 시작 하였다. 새 교육 방법으로 학생들을 교육 한 후, 학생들의 시험 성적 결과를 측정하였다. 시험 성적 결과를 바탕으로 새로이 개발된 A, B 방법은 학생들간 시험성적 차이를 만들었다고 할 수 있는가?
(단 두 그룹에 무작위로 나뉘어진 학생들의 기본 능력은 동일하다고 가정한다)

Group Statistics

	학생그룹	N	Mean	Std. Deviation	Std. Error Mean
시험결과	1.00	20	79.1000	7.9532	1.7784
	2.00	20	68.1500	9.2978	2.0791

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
시험결과	Equal variances assumed	.351	.557	4.002	38	.000	10.9500	2.7359	5.4115	16.4885
	Equal variances not assumed			4.002	37.109	.000	10.9500	2.7359	5.4071	16.4929

위의 표를 바탕으로, 두 모집단 평균차이 검증에서는 두 모집단의 분산이 같다는 가정 하에 t-test 를 실시한다. 두 모집단의 분산의 동질성 가정을 위해 Levene's test 를 사용한다. 이때 F=.351, P (Sig) =0.557)>0.005 이므로 두 모집단의 분산이 동일하다는 가정을 한다.
또한 분산이 같다는 가정 하에 (Equal variances assumed) t-value 가 4.002 이고, P (Sig, 2-tailed) = 0.000 < 0.005 으로 나타나 귀무가설 (두 방법에 차이가 없다) 은 기각된다. 다시 말해서 두 가지 교육 방법에는 차이가 있다라고 말할 수 있다. A그룹과 B그룹의 평균 차이는

10.95 이며 통계적으로 차이가 있고, 신뢰구간이 0을 포함 하지 않고 있다 (귀무가설 기각).

(3) 동일 모집단으로부터 두 표본

Analyze → Compare means → Paired-Sample T test

이 기법은 위의 두 모집단의 평균차이검증 기법과 비슷하나, 표본들의 값들이 짝을 이루고 있고, 이러한 짝 (Pair) 들에 대한 비교를 하는 경우가 있다. 단, 이러한 짝들은 서로 독립적이지 않으면 반드시 하나의 모집단에서 나온것이다. 즉, 동일집단에서 다른 변수나 동일한 변수를 여러번(일반적으로 실험효과를 보기위한 사전, 사후검사) 측정한 것을 이용한다. 이 경우는 두 변수의 데이터에 상관관계가 존재하기 때문에 평균을 비교하기 위해서 독립표본 t 검정을 이용할 수 없다.
마찬가지로 모집단이 정규본포를 따른다는 가정하에서 분석을 하기 때문에 소표본 이거나 정규본포를 따르지 않는 경우는 비모수검정인 윌콕슨(Wilcoxon)의 부호순위검정이나 부호(sign)검정을 이용하여 분석한다.

예를 들면 아래와 같다.

어느 회사에서 개발한 다이어트 신약이 환자들에게 효과가 있는지 알아보기 위해서 20여명의 지원자를 선발하였다. 20여명의 지원자를 대상으로 다이어트 신약을 사용하였고, 사용전 과 사용후 결과를 얻었다. 과연 이 다이어트 신약이 효과가 있을까?

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	사용전	62.4000	20	5.8165	1.3006
	사용후	62.7000	20	5.7225	1.2796

Paired Samples Test

	Paired Differences					t	df	ig. (2-tailed
	Mean	td. Deviatio	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 사용전 - 사용	-.3000	2.2266	.4979	-1.3421	.7421	-.603	19	.554

위의 표에 나타났듯이, 다이어트 신약을 사용하기 전 평균 몸무게와 이후의 평균 몸무게의 차이가 0.3kg 이며, 표준 편차는 2.23, 표준오차 는 0.4979 이다. 평균의 차이의 95% 신뢰구간 (-1.3421 ~ 0.7421) 이 0 을 포함하므로 다이어트 신약은 효과가 없는 것으로 나타났다. 또한 t-value 는 -0.603, P (Sig. 2-tailed) = 0.554로 나타나, 유의수준 0.05에서 두 집단간의 평균차이는 유의하지 않는 것으로 나타났다.

여기에서 t 검정통계량의 부호가 음수가 나오는 것은 t검정 통계량을 산출할 때 사용전의 평균에서 사용후의 평균값을 빼어서 산출되기 때문이며 즉 양수인 경우는 사용전의 평균이 다소 높음을 인식할 수 있고 음수인 경우는 사용후의 평균이 높음을 암시한다.

6. 일원분산분석 (One-Way ANOVA)

Analyze → Compare means → One-Way ANOVA
Dependent List: 종속변수, Factor: 독립변수,
Contrast: Polynomial
Post Hoc: Scheffe
Options: Descriptive, Homogeneity-of-variance

앞에서 두 개의 독립 모집단 평균차이 검증을 위해서 t-test 가 사용될 수 있음을 알아 보았다. 단일변량 분산분석은 (analysis of variance, ANOVA)은 세개 이상의 모집단의 평균값을 비교할 때 사용하는 기법이다. 특히 분산분석은 실험계획(experimental Design)에 의해 얻어진 데이터 분석에 많이 이용된다.
따라서 실험계획법에 따라서 달리 적용가능하지만 여기에서는 일반적인 경우(완전랜덤계획법)에 분산분석만 언급한다.

독립변수(요인변수)는 명목, 순서척도가 사용되며, 종속변수는 구간, 비율척도가 사용된다.
여기에서도 표본은 정규분포를 따르는 것을 가정하며 이 조건을 만족하지 못하는 경우는 비모수적인 방법인 크루스칼-왈리스 검정(Kruskal-Wallis)을 사용한다. 다른 가정으로 등분산의 가정이 존재한다. 이는 levene의 등분산의 검정으로 확인 가능하다.

예를 들어서, 회사 사원들의 '전공'이 회사 '총 근무년수' 와 관계 있는지 알아본다면,

Descriptives

총근무년

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
인문계	25	11.56	9.20	1.84	7.76	15.36	1	34
자연계	15	17.00	8.67	2.24	12.20	21.80	1	30
사범계	4	8.00	7.35	3.67	-3.69	19.69	2	18
예체능계	4	12.25	10.87	5.44	-5.05	29.55	1	23
Total	48	13.02	9.22	1.33	10.34	15.70	1	34

위의 표는 각 집단간 총 케이스수, 평균, 표준편차, 표준오차, 95% 신뢰구간 등을 나타내고 있다.

Test of Homogeneity of Variances

총근무년

Levene Statistic	df1	df2	Sig.
.415	3	44	.743

위의 표는 분산분석을 나타낸 표로써, 모집단의 분산이 동일한 분산을 가지고 있다는 가정하에서 실시된다. 표에는 Levene 통계량이 제시되어 있는데, Levene 통계량이 0.415 이며 P (Sig) = 0.743 > 0.005 으로써 모집단의 분산이 동일 하다는 귀무가설이 채택된다. 그러므로 다음의 분석이 가능한 것이다.

ANOVA

총근무년

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	394.069	3	131.356	1.606	.202
Within Groups	3598.910	44	81.793		
Total	3992.979	47			

위의 표에서 F = 1.606, P (Sig.) =0.202 로써 그룹간 평균이 동일하다는 귀무가설은 채택된다. 다시 말해서 회사원들의 '전공' 은 회사 '총 근무년수' 에 차이가 없다고 볼 수 있다.

그러나 독립표본 과는 달리 3개이상의 표본간의 평균차이비교를 하는 관계로 분산분석에서 통계적으로 유의한 차이가 나타난 경우에는 다중비교를 통하여 각 집단간의 유의한 차이를 구체적으로 확인을 할 필요가 있다.
이때 여러가지 방법을 이용하지만 가장 간단한 방법으로 독립표본 t 검정을 거듭실시한 효과를 나타내는 LSD(Least Significant Difference)이 있으며 그외 Bonferroni, Scheffe, Tukey, Duncan을 일반적으로 많이 이용한다. 그러나 등분산이 가정되지 않은 경우에는 다른 방법으로 다중비교를 실시해야 한다.

● Post Hoc Tests

Multiple Comparisons

Dependent Variable: 총근무년
Scheffe

(I) 전공	(J) 전공	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
인문계	자연계	-5.44	2.95	.347	-14.03	3.15
	사범계	3.56	4.87	.911	-10.60	17.72
	예체능계	-.69	4.87	.999	-14.85	13.47
자연계	인문계	5.44	2.95	.347	-3.15	14.03
	사범계	9.00	5.09	.383	-5.79	23.79
	예체능계	4.75	5.09	.832	-10.04	19.54
사범계	인문계	-3.56	4.87	.911	-17.72	10.60
	자연계	-9.00	5.09	.383	-23.79	5.79
	예체능계	-4.25	6.40	.931	-22.84	14.34
예체능계	인문계	.69	4.87	.999	-13.47	14.85
	자연계	-4.75	5.09	.832	-19.54	10.04
	사범계	4.25	6.40	.931	-14.34	22.84

위의 표는 Scheffe 방법으로 계산된 사후검증을 보여주는 것으로 유의수준 0.05 에서 전공간 유의한 차이가 있는 것을 보여주는 부분은 없다.

Homogeneous Subsets

총근무년

Scheffe[a,b]

전공	N	Subset for alpha = .05
		1
사범계	4	8.00
인문계	25	11.56
예체능계	4	12.25
자연계	15	17.00
Sig.		.364

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 6.593.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

위의 표에서는 평균차이가 없는 동일한 집단을 나타내고 있는데, 각각 4개의 전공이 모두 동일한 집단이라는 결과를 보여주고 있다 (Subset 그룹이1개).

7. 이원분산분석 (Two-Way ANOVA)

General Linear Model → Univariate
Dependent Variable: 종속변수, Fixed Factor(5): 1개 이상 독립변수
Post Hoc: Scheffe (Bonferroni, Tukey 는 옵션)

이원분산분석 (Two-Way ANOVA)이 일원분산분석 (One-Way ANOVA) 과 다른점은 독립변수가 2개 이상이란 점이다.

만약 회사 사원들의 '담당업무' 와 '대학유형' 이 회사 '총 근무년수' 와 관계를 알아본다면,

Between-Subjects Factors

		Value Label	N
담당업무	1	부서A	5
	2	부서B	36
	3	부서C	8
대학유형	1	4년제 대학	20
	2	2년제 대학	18
	3	기타(각종학교, 대학원대학 등)	11

위의 표는 담당업무 및 학력 변수에 따른 케이스들의 수를 보여주고 있다.

Tests of Between-Subjects Effects

Dependent Variable: 총근무년

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	946.633[a]	8	118.329	1.422	.217
Intercept	5418.157	1	5418.157	65.107	.000
담당업무	793.169	2	396.584	4.766	.014
대학유형	122.948	2	61.474	.739	.484
담당업무 * 대학유형	281.327	4	70.332	.845	.505
Error	3328.755	40	83.219		
Total	13031.000	49			
Corrected Total	4275.388	48			

a. R Squared = .221 (Adjusted R Squared = .066)

위의 표는 총 근무년수에 대한 담당업무 와 대학별 유형, 그리고 두 변수의 상호작용효과를 나타내고 있다.
우선 담당업무의 경우 F=4.766, P (Sig) = 0.014 < 0.05 로써 귀무가설이 기각되었다. 즉, 담당업무와 총 근무년수는 차이가 있는 것으로 나타났다.
대학유형의 경우 F=0.739, P (Sig) = 0.484 > 0.05 로써 귀무가설이 채택되었다. 즉, 대학유형과 총 근무년수는 차이가 없는 것으로 나타났다.
담당업무, 대학유형 상호작용의 경우 F=0.845, P (Sig) = 0.505 > 0.05 로써 귀무가설이 채택되었다. 즉, 상호작용과 총 근무년수는 차이가 없는것으로 나타났다.

● 사후검정 (담당업무)

Multiple Comparisons

Dependent Variable: 총근무년
Scheffe

(I) 담당업무	(J) 담당업무	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
부서A	부서B	10.71	4.35	.060	-.36	21.77
	부서C	13.27*	5.20	.049	5.34E-02	26.50
부서B	부서A	-10.71	4.35	.060	-21.77	.36
	부서C	2.57	3.57	.773	-6.50	11.63
부서C	부서A	-13.27*	5.20	.049	-26.50	-5.34E-02
	부서B	-2.57	3.57	.773	-11.63	6.50

Based on observed means.
*. The mean difference is significant at the .05 level.

위 표는 담당업무 부서간 총 근무년수에 대한 사후검증결과를 보여주고 있다. 부서A 와 부서C 가 유의적인 차이가 있는 것으로 나타났다(P = 0.049 < 0.05).
이곳에서는 Scheffe 방법만을 사용했으나 Tukey 혹은 Bonferroni 방법등도 좋은 기법중 하나라고 할 수 있다.

총근무년

Scheffe[a,b,c]

담당업무	N	Subset 1	Subset 2
부서C	8	10.13	
부서B	36	12.69	12.69
부서A	5		23.40
Sig.		.845	.065

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = 83.219.
a. Uses Harmonic Mean Sample Size = 8.504.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.
c. Alpha = .05.

* 대학 유형별 사후검정은 동일한 분석 방법이므로 생략 되었음.

위의 표에서는 통계적으로 동일한 집단을 보여주고 있다. Subset 1 에서는 부서C 와 부서B, 그리고 Subset 2 에서는 부서B 와 부서A 의 평균차이가 없다는 것을 보여주고 있다.

위의 에에서는 상호작용효과가 나타나지 않았지만 실제적으로 이원변량분석의 경우는 상호작용효과에 대한 관심을 더욱 강하게 나타낸다. 이 경우에 통계적으로 유의한 상호작용효과가 존재한다고 분석된 경우에는 연구의 목적에 따라서 위의 예에서 나타난 대학유형별 담당업무간의 차이와 담당업무별 대학유형의 차이에 대한 추가적인 분석이 동반된다.

8. 다변량분산분석 (MANOVA)

General Linear Model → Multivariate
Dependent Variables: 종속변수, Fixed Factor(5): 1개 이상 독립변수
Covariate(s): 통제변수
Post Hoc: Scheffe (Bonferroni, Tukey 는 옵션)

다변량분산분석 (Multivariate Analysis of Variance: MANOVA) 는 종속변수의 수가 2 개 이상인 경우 집단 평균벡터(vector) 비교하는데 이용하는 통계 기법이라 할 수 있다. SPSS에서는 공분산분석(ANCOVA), 반복측정분산분석(Repeated Measure ANOVA)이 많이 이용된다.
그래서 지금까지의 분석들의 특징을 보면, 두 집단간 평균의 비교는 t-test, 두 집단 이상의 평균을 비교하는 데는 ANOVA, 그리고 종속변수가 2개 이상인 경우는 MANOVA 를 사용하게 된다.

그렇다면 왜 MANOVA 를 사용하나?

(1) 만약 종속변수간 강한 상관관계가 있을 경우, ANOVA를 이용해서는 밝혀 낼 수 없는 결합차이들을 알 수 있다.
(2) ANOVA 를 사용했을 경우에 비해 오차 (1종 오차)의 확률을 줄일 수 있다.

또한 MANOVA 의 가정을 보면

(1) 관측치 들이 서로 독립적 이여야 되고
(2) 변수들이 다변량 정규분포성 (Multivariate normal distribution) 을 충족해야 하며
(3) 각 집단간 분산-공분산 행렬이 동일해야 한다.

예를 들어서 회사 사원들의 ‘대학유형’ 이 회사 ‘총 근무년수’ 와 ‘월 수입’ (2개의 종속변수) 관계 있는지 알아본다면,

General Linear Model

Between-Subjects Factors

		Value Label	N
대학유형	1	4년제 대학	20
	2	2년제 대학	18
	3	기타(각종학교, 대학원대학 등)	11

위의 표는 대학 유형별 표본의 수를 나타내고 있다.

Multivariate Tests [c]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.775	77.493[a]	2.000	45.000	.000
	Wilks' Lambda	.225	77.493[a]	2.000	45.000	.000
	Hotelling's Trace	3.444	77.493[a]	2.000	45.000	.000
	Roy's Largest Root	3.444	77.493[a]	2.000	45.000	.000
대학유형	Pillai's Trace	.010	.120	4.000	92.000	.975
	Wilks' Lambda	.990	.118[a]	4.000	90.000	.976
	Hotelling's Trace	.011	.116	4.000	88.000	.977
	Roy's Largest Root	.010	.234[b]	2.000	46.000	.792

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+대학유형

위의 표는 대학유형의 다변량검증에 대한 결과를 보여주고 있다.

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	총근무년	11.931[a]	2	5.966	.064	.938
	수입	202.375[b]	2	101.187	.008	.992
Intercept	총근무년	8045.952	1	8045.952	86.811	.000
	수입	1884957.813	1	1884957.813	153.370	.000
대학유형	총근무년	11.931	2	5.966	.064	.938
	수입	202.375	2	101.187	.008	.992
Error	총근무년	4263.457	46	92.684		
	수입	565352.727	46	12290.277		
Total	총근무년	13031.000	49			
	수입	2577900.000	49			
Corrected Total	총근무년	4275.388	48			
	수입	565555.102	48			

a. R Squared = .003 (Adjusted R Squared = -.041)

b. R Squared = .000 (Adjusted R Squared = -.043)

위의 표에서는 대학 유형에 대한 총근무년수 와 수입에 대한 결과가 제시 되어있다. 우선 대학유형에 대한 총 근무년수 의 경우 F=0.064, P (Sig) = 0.938 > 0.005 이므로, 총 근무년수는 대학유형별 (2년제, 4년제, 기타) 로 차이가 있다고 할 수 없다.

또한, 우선 대학유형에 대한 수입 의 경우 F=0.008, P (Sig) = 0.992 > 0.005 이므로, 수입은 대학유형별 (2년제, 4년제, 기타) 로 차이가 있다고 할 수 없다.

- Post Hoc Tests

대학유형

Multiple Comparisons

Scheffe

Dependent Variabl	(I) 대학유형	(J) 대학유형	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
총근무년	4년제 대학	2년제 대학	.68	3.13	.977	-7.23	8.59
		기타(각종학교, 대학원 대학 등	1.26	3.61	.941	-7.88	10.41
	2년제 대학	4년제 대학	-.68	3.13	.977	-8.59	7.23
		기타(각종학교, 대학원 대학 등	.59	3.68	.987	-8.73	9.91
	기타(각종학교, 대학원 대학 등	4년제 대학	-1.26	3.61	.941	-10.41	7.88
		2년제 대학	-.59	3.68	.987	-9.91	8.73
수입	4년제 대학	2년제 대학	4.0000	36.0181	.994	-87.1136	95.1136
		기타(각종학교, 대학원 대학 등	-.5455	41.6150	1.000	-105.8172	104.7263
	2년제 대학	4년제 대학	-4.0000	36.0181	.994	-95.1136	87.1136
		기타(각종학교, 대학원 대학 등	-4.5455	42.4275	.994	-111.8724	102.7815
	기타(각종학교, 대학원 대학 등	4년제 대학	.5455	41.6150	1.000	-104.7263	105.8172
		2년제 대학	4.5455	42.4275	.994	-102.7815	111.8724

Based on observed means.

위의 표는 각 종속변수별로 집단간의 차이의 유무를 사후검증한 결과 표이다.
예를 들어서 총근무년수 에서 4년제와 2년제 간에 유의한 차이가 없는것으로 나타내고 있다.
p= .977 > 0.005). 마친가지로 모든 집단에서 유의한 차이가 없는것으로 나타나고 있다.
결과적으로 대학유형과 회사내 총근무년수 그리고 수입간에는 차이가 없는것으로 나타났다.

총근무년

Scheffe[a,b,c]

대학유형	N	Subset 1
기타(각종학교, 대학원 대학 등)	11	12.64
2년제 대학	18	13.22
4년제 대학	20	13.90
Sig.		.936

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = 92.684.

a. Uses Harmonic Mean Sample Size = 15.270.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

c. Alpha = .05.

위의 표에서는 총근무년에 대한 대학유형이 통계적으로 동일한 집단을 보여주고 있다.

- 대학유형에 대한 수입 부분 역시 같은 내용으로써 본 내용에서 삭제됨

9. 상관분석 (Correlation Analysis)

Analyze → Correlate → Bivariate

상관분석은 두 변수가 어느 정도 밀접한 관계를 가지고 있는 분석인지를 나타내주는 기법이다.

예를 들어서, 고등학교 학생들의 키와 체중 사이의 관계를 알고 싶어 하는 경우 상관분석이 사용되어야 한다. 키와 체중은 서로간 독립적이기 보다 서로 영향을 주고 받는 관계 이다. 키가 크기 때문에 체중이 많이 나간다고 할 수 있지만, 역으로 체중이 많이 나가는 학생들의 키가 또한 크기 때문이다 (다시 말해서 회귀분석의 경우 처럼 한 변수가 다른 변수에 영향을 주기 보다 서로 밀접한 상호작용을 하는것이다.)

상관관계의 종류

상관관계에는 세가지 종류로 구분될수 있다.

(1) 단순상관계수 (Simple correlation coefficient): 두 변수사이의 상관관계를 의미한다.
(2) 다중상관관계 (Multiple correlation): 하나의 변수와 다른 두개 변수 이상의 변수사이의 상관관계를 의미한다.
(3) 부분상관관계 (Partial correlation): 다른 변수의 상관관계를 통제한 상태에서, 두 변수간의 순수한 상관관계를 의미한다.

상관계수

상관계수 (correlation coefficient) 는 두 변수 사이에 관계가 얼마나 강한지 혹은 약한지를 나타내주는 지수가 되겠다. 우선 + 숫자는 양의 방향의 관계를 나타내며, - 숫자는 음의 방향의 관계를 나타낸다. 또한 지수에 대한 해석은 아래와 같다.

1.0 ~ 0.7: 매우 강한 양의 관련성, -1.0 ~ -0.7: 매우 강한 음의 관련성
0.7 ~ 0.4: 강한 양의 관련성, -0.7 ~ -0.4: 강한 음의 관련성
0.4 ~ 0.2: 약한 양의 관련성, -0.4 ~ -0.2: 약한 음의 관련성
0.2 ~ 0.0: 거의 관련이 없음, -0.2 ~ -0.0: 거의 관련이 없음

우선 조사한 고등학생의 키와 체중 그리고 나이에 관한 상관관계를 조사한다고 가정할 경우,

단순 상관분석 결과

Correlations

		키	체중	나이
키	Pearson Correlation	1.000	.763**	.530**
	Sig. (2-tailed)	.	.000	.000
	N	50	50	50
체중	Pearson Correlation	.763**	1.000	.603**
	Sig. (2-tailed)	.000	.	.000
	N	50	50	50
나이	Pearson Correlation	.530**	.603**	1.000
	Sig. (2-tailed)	.000	.000	.
	N	50	50	50

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

위에 표에서 보는 바와 같이, 키와 체중의 상관계수는 0.763 으로써, 강한 양의 상관관계 보여주고 있으며, 통계적으로도 매우 유의한 것으로 나타났다 유의수준 .001). 또한 키와 나이의 경우 0.530, 나이와 체중의 경우 0.6 으로써, 나이 역시 키와 체중에 상당한 상관관계가 있음을 보여주고 있다.

10. 부분 상관분석

Analyze → Correlate → Partial

만약 조사자가 학생들의 나이를 통제하고, 순수하게 키와 체중만의 상관관계를 알고 싶어할 경우, 다음과 같은 결과를 얻을 수 있다.

부분상관분석결과

```
- - -  P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S  - - -
Controlling for..     나이
                   키          체중
키              1.0000       .6558
              (     0)     (    47)
              P= .         P= .000
체중             .6558      1.0000
              (    47)     (     0)
              P= .000      P= .
(Coefficient / (D.F.) / 2-tailed Significance)
" . " is printed if a coefficient cannot be computed
```

위에서 보는바와 같이, 나이를 통제한 상태에서 키와 체중의 상관관계는 0.6558 로써 강한 상관관계를 나타내고 있다. 또한 통계적으로도 매우 유의한 관계 (P = .000)를 보여주고 있다.

11. 회귀분석 (Regression)

Analyze → Regression → Linear
Dependent: 종속변수, Independent: 독립변수

회귀분석 (Regression)은 독립변수와 종속변수간 인과관계(Causal relationship)를 알아보는 기법이라고 할 수 있겠다. 독립변수(Independent Variable)는 다른 변수에 영향을 주는 변수이며, 종속변수 (Dependent Variable)는 이와 반대로 영향을 독립변수에 영향을 받는 변수이다.

회귀분석은 단순회귀분석 (Simple regression analysis) 와 중회귀분석 (Multiple regression analysis) 으로 크게 분류되는데, 단순회귀분석의 경우는 독립변수와 종속변수가 각기 하나씩인 경우이며, 중회귀분석은 여러 개의 독립변수와 하나의 종속 변수가 사용된 경우이다. 결국 회귀분석과 중회귀 분석의 차이점은 독립변수의 수에 그 차이가 있다고 할 수 있다.

(1) 단순회귀분석 (Simple regression analysis)

예를 들어서, 조사자가 신상품에 대한 만족도를 조사 하려고 한다고 가정하자. 이럴경우 신상품에 대한 브랜드인지도, 디자인, 가격, 성능, 품질이 전반적인 만족에 영향을 미치는지 알아보기 위해서 회귀분석을 실시 하였다.

위의 경우에서는 브랜드인지도, 디자인, 가격, 성능, 품질이 독립변수가 되며, 전반적인 만족이 종속변수가 되겠다.
설문문항 예제 가 아래와 같을 경우

	설문문항
브랜드 인지도	본 상품의 브랜드에 대해서 어느 정도 알고 계십니까? (1) 전혀 모른다 (2) 잘모르겠다 (3) 그저그렇다 (4) 잘안다 (5) 매우 잘안다
디자인	본 상품의 디자인에 대해서 어떻게 생각하십니까? (1) 전혀 맘에 들지 않는다 (2) 맘에 들지 않는다 (3) 그저그렇다 (4) 맘에 든다 (5) 매우 맘에 든다.
가격	본 상품의 가격에 대하여 어떻게 생각 하십니까? (1) 매우 비싸다 (2) 비싸다 (3) 그저그렇다 (4) 비싸지 않다 (5) 전혀 비싸지 않다
성능	본 신상품의 성능 (Function) 에 대해서 어떻게 생각 하십니까? (1) 전혀 맘에 들지 않는다 (2) 맘에 들지 않는다 (3) 그저그렇다 (4) 맘에 든다 (5) 매우 맘에 든다.
품질	본 신상품의 품질 대해서 어떻게 생각 하십니까? (1) 전혀 맘에 들지 않는다 (2) 맘에 들지 않는다 (3) 그저그렇다 (4) 맘에 든다 (5) 매우 맘에 든다.
전반적만족	신상품에 대하여 귀하께서는 어느정도 만족 하고 계십니까? (1) 매우 불 만족한다 (2) 불만족 한다 (3) 그저그렇다 (4) 만족한다 (5) 매우 만족한다

단순회귀분석의 경우, 위에서 잠깐 언급했듯이 하나의 독립과 하나의 종속변수로 구성되어 있다. 그래서 브랜드와 전반적인 만족의 관계를 알아보았다.

단순회귀분석 결과표

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.346[a]	.120	.101	.86

a. Predictors: (Constant), 브랜드

위의 표에서는 브랜드 인지도 (브랜드) 가 독립변수, 전반적 만족 (만족)이 종속변수 임을 나타내고 있다.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	브랜드[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: 만족

위의 표에서는 R Square (R^2)값을 보여주고 있다. R^2 (결정계수) 는 회귀선에 의해서 설명이 되어지는 비율을 나타내는 것으로, 브랜드에 의해서 신상품의 전반적인 만족의 12.0% 설명이 되어진다는 것을 보여주고 있다.
R^2 의 범위는 0 과 1 사이의 값을 가지게 된다. R^2 값이 1 에 가까운 수록 회귀선을 설명하는데 유용하다고 할 수 있다.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.805	1	4.805	6.520	.014[a]
	Residual	35.375	48	.737		
	Total	40.180	49			

a. Predictors: (Constant), 브랜드

b. Dependent Variable: 만족

위의 표는 회귀식이 얼마나 통계적으로 유의한지 검정하는 표로써, Sig = 0.014 < 0.05 이므로 본 회귀식이 통계적으로 유의 하다고 할 수 있다 (일반적으로 Sig 가 0.005 보다 작을 경우 통계적으로 유의 하다고 본다).

Coefficients [a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2.263	.530		4.268	.000
	브랜드	.388	.152	.346	2.553	.014

a. Dependent Variable: 만족

위의 표는 회귀계수를 보여주는 중요한 표로써, 회귀식의 상수값은 2.263 이고, Sig = .000 < .005 로써 통계적으로 유의함을 보여주고 있다.
또한 브랜드 인식도에 대한 회귀계수는 0.388 이며 Sig = .014 < .005 이므로 통계적으로 유의하다 고 할 수 있다. 본 테스트에 대한 회귀식은 다음과 같다.

Y = 2.263 (상수) + 0.388 X

Y = 전반적만족도
X = 브랜드 인지도

결론적으로, 브랜드 인지도가 통계적으로 전반적인 소비자 만족도에 영향을 미치지만 그 영향력은 매우 작다고 할 수 있다 ($R^2 = 0.120$)

(2) 중회귀분석 (Multiple regression analysis)

중회귀분석은 다수의 독립변수가 하나의 종속변수에 미치는 영향을 나타내는 것으로, 단순회귀분석(Simple regression analysis) 보여준 것과 달리 독립변수에 브랜드 인지도 외에, 디자인, 가격, 기능, 품질 모두가 전반적인 만족에 영향을 미치는지 알아보기 위해서 실시 되는 분석이라고 할 수 있겠다.

Variables Entered/Removed [b]

Model	Variables Entered	Variables Removed	Method
1	기능, 브랜드, 디자인, 품질, 가격[a]	.	Enter

a. All requested variables entered.
b. Dependent Variable: 만족

위의 표에서는 기능, 브랜드, 디자인, 품질, 가격이 독립변수, 전반적 만족 (만족)이 종속변수임을 나타내고 있다.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.790ᵃ	.625	.582	.59

a. Predictors: (Constant), 기능, 브랜드, 디자인, 품질, 격

위의 표에서는 R Square (R^2)값을 보여주고 있다. R^2 (결정계수) 는 회귀선에 의해서 설명이 되어지는 비율을 나타나는 것으로, 기능, 브랜드, 디자인, 품질, 가격에 의해서 신상품의 전반적인 만족의 62.5% 설명이 되어진다는 것을 보여주고 있다. 단순회귀분석에서 브랜드 만에 의해서 설명되어지는 설명량보다 훨씬 높음을 보여주고 있다.

ANOVA [b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	25.105	5	5.021	14.655	.000ᵃ
	Residual	15.075	44	.343		
	Total	40.180	49			

a. Predictors: (Constant), 기능, 브랜드, 디자인, 품질, 가격

b. Dependent Variable: 만족

위의 표는 회귀식이 얼마나 통계적으로 유의한지 검정하는 표로써, Sig = 0.000 < 0.05 이므로 본 회귀식이 통계적으로 유의 하다고 할 수 있다

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.372	.518		-.718	.477
	브랜드	.238	.117	.212	2.042	.047
	가격	.106	.109	.112	.971	.337
	디자인	.107	.097	.110	1.105	.275
	품질	.304	.111	.312	2.747	.009
	기능	.393	.120	.401	3.270	.002

a. Dependent Variable: 만족

위의 표에서 알 수 있듯이, 브랜드 인식도에 대한 회귀계수는 0.047 이며 Sig = .014 < .005 이므로 통계적으로 유의하다. 이러한 방법으로 품질 및 기능이 통계적으로 유의함을 알 수 있다.

따라서 전반적인 만족도는 기능에 대한 만족도가 가장 강한 영향력을 주며 다음으로 품질, 브랜드이미지의 순으로 향상효과가 있는 것으로 나타났다.

본 테스트에 대한 다중 회귀식은 다음과 같다.

$Y = -0.372$ (상수) $+ 0.238 X_1 + 0.106 X_2 + 0.107 X_3 + 0.304 X_4 + 0.393 X_5$

Y = 전반적만족도
X_1 = 브랜드
X_2 = 가격
X_3 = 디자인
X_4 = 품질
X_5 = 기능

그러나 이모형에서는 전반적인 만족도에 유의한 영향력을 미치지 못하는 변인 또한 독립변인으로 포함하여 분석한 모형이므로 최종적인 모형을 산출하기 위해서는 단계적인 방법(Stepwise method)나 전진 선택법, 후진 선택법 등을 이용하여 최종 모형을 산출할 수 있다.

13. 신뢰성 분석 (Reliability)

Analyze → Scale → Reliability
Statistics → Descriptives for: Item, Scale, Scale if item deleted
Statistics → Inter item: corrleation

신뢰성은 동일한 개념을 대상으로 반복적인 측정을 했을 경우 나타나는 측정값들의 일관된 정도를 나타낸다.

신뢰성을 측정하는 방법에는 재측정 신뢰도 (test-retest reliability), 반분 신뢰도 (split-half reliability), 문항분석 (item-total correlation), Chronbach alpha 등 여러가지 기법들이 있다.

우선 기본적으로 신뢰성 분석은 측정문항이 2개 이상 이여야 한다.

예를 들어서, 회사내 사원들의 '자율성'를 측정하기위해서 5문항이 측정되었는데, 이 문항들의 신뢰성을 알아보기로 했다. 결과는 다음과 같다.

R E L I A B I L I T Y   A N A L Y S I S - S C A L E (A L P H A)

		Mean	Std Dev	Cases
1.	자율1	3.6400	.7217	50.0
2.	자율2	3.1800	.8003	50.0
3.	자율3	3.4000	.7284	50.0
4.	자율4	3.6400	.9424	50.0
5.	자율5	3.4000	.7284	50.0

→ 위의 표는 자율문항의 평균, 표준편차, 그리고 케이스 개수를 보여주고 있다.

Correlation Matrix

	자율1	자율2	자율3	자율4	자율5
자율1	1.0000				
자율2	.7506	1.0000			
자율3	.5513	.5041	1.0000		
자율4	.4957	.5206	.0654	1.0000	
자율5	.5901	.6092	.6538	.2438	1.0000

→ 위의 표는 자율문항 간 상관관계를 보여주고 있다. 자율성을 측정한 1번문항과 2번 문항의 상관관계 (0.7506) 가 높은 것으로 나타나고 있다.

N of Cases = 50.0

Statistics for	Mean	Variance	Std Dev	N of Variables
Scale	17.2600	9.0535	3.0089	5

→ 위의 표는 5문항을 하나로 가정한 상황에서 평균, 분산, 표준편차, 문항의 수를 보여주고 있다.

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
자율1	13.6200	5.7914	.7892	.6420	.7371
자율2	14.0800	5.4629	.7886	.6446	.7304
자율3	13.8600	6.5718	.5224	.5201	.8099
자율4	13.6200	6.2812	.3988	.3803	.8616
자율5	13.8600	6.1637	.6523	.5392	.7752

→ 위의 표에서 가장 중요한 부분은 'Alpha if item deleted' 로써 각 문항을 제거 했다고 가정했을 경우 Chronbach a 값이 되겠다. 다시 말해서 수치가 높으면 높을수록 그 문항을 제거 할 경우 Chronbach a 값이 높아지는 것이다. 예를 들어서 자율4번을 제거하면 Chronbach a 값은 0.8616 으로 올라가고, 반대로 자율2 번을 제거하면 Chronbach a 값은 0.7304 로 떨어지게 된다.

Reliability Coefficients 5 items
Alpha = .8205 Standardized item alpha = .8325

→ 위의 표를 보아서 자율 5문항의 Chronbach a (Alpha) 값은 0.8205 이며, 표준화된 각 문항으로 신뢰성 분석을 했을 경우 Chronbach a (Standardized item alpha) 값은 0.8325 임을 알 수 있다.
일반적으로 Alpha 의 사용이 보편화되고 있으나, 문항의 척도가 큰 분산을 이루고 있는 경우 Standardized item alpha 를 사용해야 된다.

일반적으로 cronbach s Alpha의 값이 0.6이상이면 신뢰성이 높은 것으로 판단한다.

14. 요인분석 (Factor Analysis)

Analyze → Data Reduction → Factor
Variables: 측정항목 변수
Rotation: Varimax
Extraction: Scree
Scores: Save as variables (요인을 변수로 만들고 싶을 때)

요인분석은 여러 측정문항들의 상관관계를 바탕으로, 측정문항의 공통차원 (common underlying dimension) 을 통해 문항의 수보다 적은 요인 (factor) 을 만들어 내는 통계기법 이다.

요인분석을 실시 하기 위해서는 문항이 간격척도나 혹은 비율척도 이여야 하며, 표본의 크기는 최소 50 개에서 100 개 정도는 되어야 한다.

요인추출 방법에는 주로 주성분분석 (Principle Component Analysis) 와 공통요인분석 (Common Factor Analysis) 등이 사용되지만, 보편적으로 주성분분석이 주로 많이 사용된다고 할 수 있다.

요인의 수 결정에 관해서는 여러 가지 방법이 사용되는데

(1) 고유치 (Eigenvalues) 기준: Eigenvalue 는 요인적재값의 제곱의 합을 나타내는데, 고유치가 1 이상인 경우를 기준으로 요인수를 결정한다.

(2) 요인수의 사전 결정: 이 방법은 요인분석전, 요인의 수를 미리 조사자가 결정하는 방법으로써, 원하는 수의 요인을 얻을 수 있다.

(3) 스크리 검정 (Scree test): SPSS 에서는 스크리 도표 (Scree table) 을 제공하다. 세로축은 각 요인의 고유치를 나타내고 가로축은 요인의 개수를 나타내는데, 감소폭이 갑자기 줄어드는 부분의 전까지 요인의 수를 기준으로 요인을 추출 할 수 있다.

(4) 공통분산의 총분산에 대한 비율: 이 방법은 요인들의 설명력의 합을 사전에 어느 정도 결정한 후 지정하는 방법으로써, 보통 사회과학분야에서는 60% 정도를 기준으로 삼고 있다.

어떤 문항이 특정 요인에 높게 관계되는지 알아보기 위해서 요인을 회전하게 된다. 회전방식에는 직각회전방식 (Varimax, Quartimax, Equimax) 과 사각회전 (Oblimin, Promax) 등이 있는데 Varimax 방식이 가장 보편적인 방법이다.

예를 들어서 회사낸 직원들에 대한 만족도를 측정하기 위해서 여러 가지 문항(X1~X9) 을 만들어서 조사한 후 요인들 알아보기 위해 실시한 요인분석 결과는 아래와 같다
(Options: Coefficient Display Format: suppress → 0.50)

Factor Analysis

Communalities

	Initial	Extraction
X1	1.000	.720
X2	1.000	.690
X3	1.000	.580
X4	1.000	.719
X5	1.000	.637
X6	1.000	.547
X7	1.000	.720
X8	1.000	.470
X9	1.000	.702

Extraction Method: Principal Component Analysis.

위의 표는 공통성 (Communalities)을 보여주고 있는데, 공통성은 0 과 1 사이에 값을 갖는다. 예를 들어서 X1 의 경우 추출된 요인들에 의해서 72% 가 설명되고 있음을 알 수 있다.

Total Variance Explained

	Initial Eigenvalues			action Sums of Squared Loa			ation Sums of Squared Loadi		
Compon	Total	of Varian	umulative	Total	of Varian	umulative	Total	of Varian	umulative
1	4.555	50.610	50.610	4.555	50.610	50.610	3.274	36.382	36.382
2	1.230	13.670	64.280	1.230	13.670	64.280	2.511	27.898	64.280
3	.978	10.863	75.143						
4	.619	6.879	82.022						
5	.421	4.675	86.696						
6	.409	4.547	91.243						
7	.314	3.490	94.733						
8	.259	2.880	97.613						
9	.215	2.387	100.000						

Extraction Method: Principal Component Analysis.

위의 표는 9개 문항이 9개 요인으로 추출되는 고유치 (Eigenvalue) 와 설명력을 보여주고 있다. 고유치 1 을 이용하였을 경우, 요인 3의 고유치가 0.979 (Total) 이므로, 요인 2개만이 추출이 된다.

Component Matrix [a]

	Component	
	1	2
X1	.626	.573
X2	.813	
X3	.541	.536
X4	.780	
X5	.751	
X6	.730	
X7	.764	
X8	.592	
X9	.754	

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

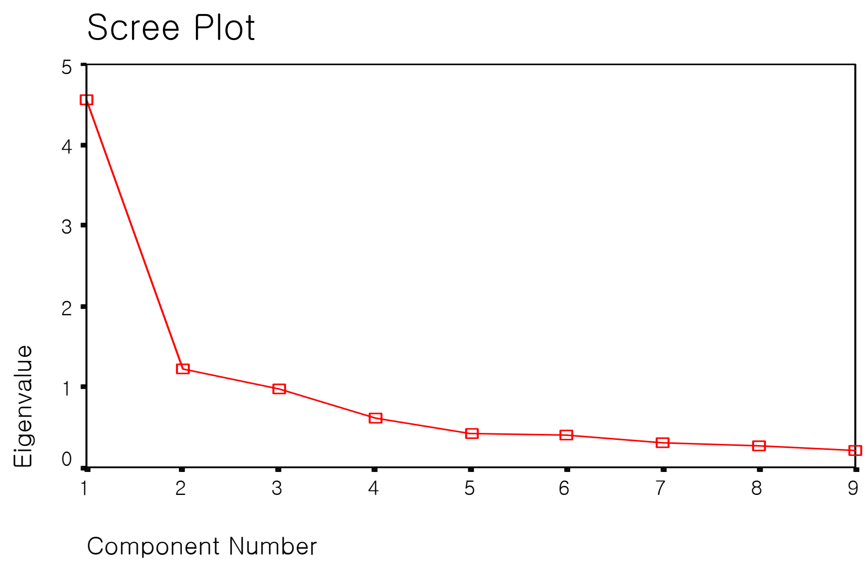
위의 표는 요인행렬이 회전되기 전의 요인 부하량을 보여주고 있다.

Rotated Component Matrix [a]

	Component	
	1	2
X1		.838
X2	.529	.641
X3		.756
X4		.745
X5	.756	
X6	.646	
X7	.828	
X8	.678	
X9	.818	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

위의 표는 회전 후의 요인 부하량을 나타내고 있다. 요인 회전 전과 많은 차이를 보여주고 있는데, 요인 1에는 X5, X6, X7, X8, X9 그리고 요인 2 에는 X1, X2, X3, X4 의 요인이 높게 걸려있다. X2 의 경우 두 요인 모두 0.5 이상의 요인부하량을 보여주고 있으나 요인2에 좀더 높은 요인 부하량을 보여주고 있다.



위의 표는 스크리 도표로써, 가로축이 요인수, 그리고 세로축이 고유치가 된다.

Component Transformation Matrix

Component	1	2
1	.784	.621
2	-.621	.784

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

위의 값은 요인회전시 사용된 성분변환 행렬값을 보여주고 있다.

15. 판별분석 (Discriminant Analysis)

판별분석은 한 개체가 어느 집단에 속하는가를 예측하는 분석으로 한개 이상의 설명(독립)변수를 가지고 집단(종속)변수에 대한 판별모형을 만든다.
예를 들면, 은행에서 신용도를 가지고 도산위험이 높은 집단과 낮은 집단으로 분류된다. 즉, 기존에 있는 데이터를 바탕으로 관찰되어지는 개체들을 몇 개의 그룹으로 분류하는 경우 사용된다. 특히 독립변수는 등간척도나 비율척도가 사용되며 종속변수로는 명목, 순서척도가 사용된다. 그러나 범주형 자료를 이용하는 경우에는 반드시 더미변수로 변환하여 이용하여야 한다.

다음의 자료는 Fisher의 붓꽃자료이며 세가지 품종(setosa, Versicolor, Virginica)의 붓꽃으로부터 각각 50개씩 추출하여 꽃받침조각의 길이(x1), 꽃받침조각의 폭(x2), 꽃잎의 길이(x3), 꽃잎의 폭(x4)을 각각 측정한 자료이다.

통계분석-> 분류분석->판별분석
집단변수-> variety, 독립변수-> x1, x2, x3, x4
통계량: 평균, 일변량분산분석, Box의 M, Fisher의 방법, 표존화 하지 않음
분류 : 결합-진단, 영역도

분석 결과는 다음과 같다.

먼저 집단변수의 각 집단(품종)에 대한 독립변수들의 기술통계량이며 집단평균의 동질성에 대한 검정을 보면 p=.000으로 네가지 품종에 대해서 4개의 측정치에 차이가 있는 것으로 나타났다. 즉 네가지 독립변수는 품종을 판별할 수 있는 판별변수로서 적합하다고 판단할 수 있다.

집단 통계량

품종		평균	표준편차	유효수(목록별) 가중되지 않음	유효수(목록별) 가중됨
setosa	sepal length	5.006	.3525	50	50.000
	sepal width	3.428	.3791	50	50.000
	petal length	1.462	.1737	50	50.000
	petal width	.246	.1054	50	50.000
versicolor	sepal length	5.936	.5162	50	50.000
	sepal width	2.770	.3138	50	50.000
	petal length	4.260	.4699	50	50.000
	petal width	1.326	.1978	50	50.000
virginica	sepal length	6.570	.6102	50	50.000
	sepal width	2.974	.3225	50	50.000
	petal length	5.552	.5519	50	50.000
	petal width	2.026	.2747	50	50.000
합계	sepal length	5.837	.8162	150	150.000
	sepal width	3.057	.4359	150	150.000
	petal length	3.758	1.7653	150	150.000
	petal width	1.199	.7622	150	150.000

집단평균의 동질성에 대한 검정

	Wilks 람다	F	자유도1	자유도2	유의확률
sepal length	.377	121.652	2	147	.000
sepal width	.599	49.160	2	147	.000
petal length	.059	1180.161	2	147	.000
petal width	.071	960.007	2	147	.000

BOX' M은 집단간 공분산행렬이 동일성에 대한 검정 결과 p=.000 으로 집단간 공분산행렬이 동일하다고 볼 수 없기 때문에 분류방법 적용시 개별집단 공분산 행렬을 사용하여야 한다.

공분산행렬의 동일성에 대한 Box의 검정

로그 행렬식

품종	순위	로그 행렬식
setosa	4	-13.067
versicolor	4	-10.874
virginica	4	-8.884
집단-내 통합값	4	-9.930

인쇄된 판별값의 순위와 자연로그는 집단 공분산행렬의 순위 및 자연로그를 나타냅니다.

검정 결과

Box의 M		148.697
F	근사법	7.143
	자유도1	20
	자유도2	77566.751
	유의확률	.000

모집단 공분산행렬이 동일하다는 영가설을 검정합니다.

아래의 결과에서는 판별분석에서 구한 2개의 정준판별함수(canonical discriminant function)에 대한 표로 함수1의 고유값은 32.325이고, 전체 판별력의 99.1%를 설명하며, Wilks람다값과 카이제곱의 유의확률이 0.000이므로 함수1의 판별력과 함수2의 판별력을 추가함으로써 판별력이 모두 유의하다고 할 수 있다.

정준 판별함수의 요약

고유값

함수	고유값	분산의 %	누적 %	정준 상관
1	32.325[a]	99.1	99.1	.985
2	.285[a]	.9	100.0	.471

a. 첫 번째 2 정준 판별함수가 분석에 사용되었습니다.

Wilks의 람다

함수의 검정	Wilks의 람다	카이제곱	자유도	유의확률
1에서 2	.023	546.697	8	.000
2	.778	36.530	3	.000

다음의 표준화 정준 판별함수계수와 각 독립변수와 판별함수의 통합(pooled)상관계수이다. 즉, 판별함수1은 꽃잎의 길이와 가장 상관성이 높으며, 다음으로 꽃잎의 폭이며, 판별함수2는 꽃받침조각의 폭과 가장 상관이 높으며 다음으로 꽃잎의 폭이다.

표준화 정준 판별함수 계수

	함수	
	1	2
sepal length	-.425	-.014
sepal width	-.532	.746
petal length	.940	-.386
petal width	.586	.575

구조행렬

	함수	
	1	2
petal length	.705*	.165
sepal width	-.119	.864*
petal width	.632	.734*
sepal length	.225	.294*

판별변수와 표준화 정준 판별함수 간의 집단-내 통합 상관행렬. 변수는 함수내 상관행렬의 절대값 크기순으로 정렬되어 있습니다.

*. 각 변수와 임의의 판별함수 간의 가장 큰 절대 상관행렬

이때 정준판별함수의 계수를 통하여 판별점수를 산출하면,

D1=-1.933-0.842*x1-1.566*x2+2.183*x3+2.862*x4
D2=-6.560-0.027*x1+2.197*x2-0.897*x3+2.812*x4

이며 이 경우 각 집단에서 독립변수 자리에 독립변수의 중심값을 넣고 계산한 판별점수의 값이 집단 중심점이다. 즉, setosa품종은 함수1의 값이 -7.6에 가깝다는 것을 의미한다.

정준 판별함수 계수

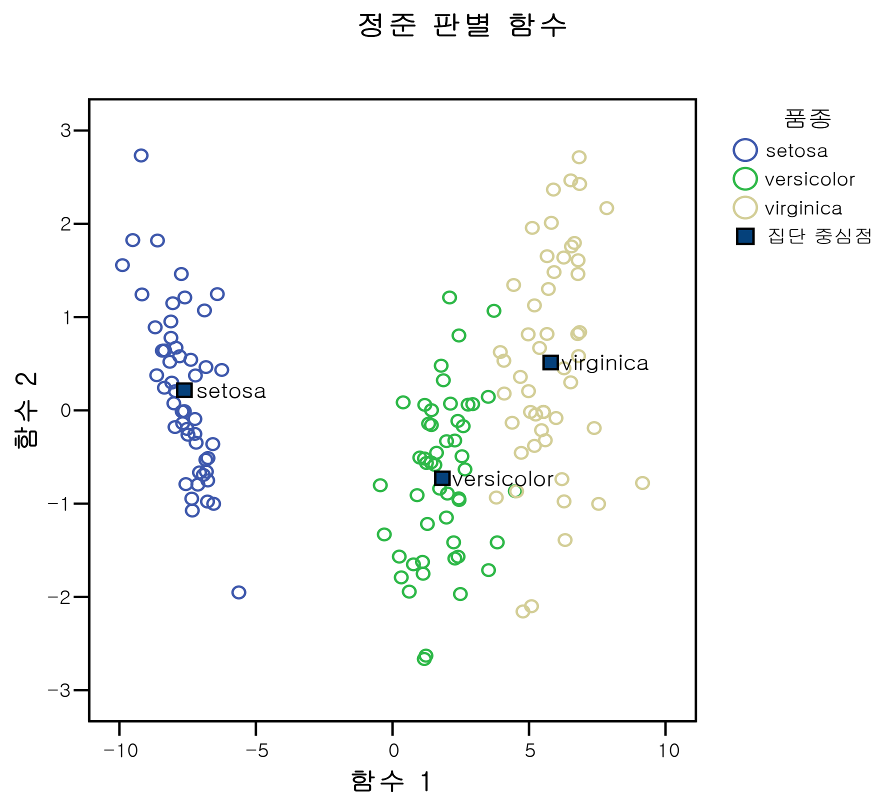
	함수	
	1	2
sepal length	-.842	-.027
sepal width	-1.566	2.197
petal length	2.183	-.897
petal width	2.862	2.812
(상수)	-1.933	-6.560

표준화하지 않은 계수

함수의 집단중심점

	함수	
품종	1	2
setosa	-7.622	.215
versicolor	1.826	-.728
virginica	5.797	.513

표준화하지 않은 정준 판별함수가 집단 평균에 대해 계산되었습니다.



16. 군집분석

군집분석은 실험의 결과나 표본에서 얻어진 대상들을 다양한 특성의 유사성을 바탕으로 동질적인 몇 개의 군집(cluster)으로 분류하는 방법이다.

예를 들어 심리학에서 성격유형에 따라서 몇개의 그룹으로 구분하거나 의학에서 환자의 증상에 따라서 몇개의 그룹으로 구분하는 방법이 이에 해당한다.
군집의 유형은 크게 계층적 군집과 상호배반적 군집으로 나누며, 계층적 군집은 한 군집이 다른 군집의 내부에 포함되나 군집간에는 중복이 허용하지 않는 형식을 취하며, 상호배반적 군집은 각 개체가 상호배반적인 여러 군집들 중 하나에만 속하는 유형이다.

각 케이스의 유사성을 측정하기 위해 거리행렬을 이용하여 상대적인 거리가 가까운 개체들 끼리 같은 군집을 이루게 하는 방법으로 유클리드 거리와 유클리드 제곱거리, 체비쉐프 거리 등을 이용한다.

(1) 계층적 군집분석

다음의 예는 9개국의 경제상태에 따른 군집분석 자료로 계층적으로 분석한 경우에는 다음과 같은 결과가 나타난다.

분석-> 분류분석-> 계층적 군집분석
변수 : gnp, trade, defense, gdp
케이스 설명 기준변수 : nation(문자타입)
통계량 : 군집화 일정표, 해법범위(3에서 5)
도표 : 덴드로그램
방법: 군집방법-가장 가까운 항목(최단연결법), 측도의 간격(제곱유클리디안 거리), 값변환(Z 점수)

계층적 군집분석을 실시한 경우에 군집을 3개로 추출할 경우에는 (1~7) (8) (9)로 분류되며, 군집을 4개로 하는 경우에는 (1, 3, 6) (2, 4, 5, 7) (8) (9)로 분류되고, 군집을 5개로 분류하는 경우에는 (1) (2, 4, 5, 7) (3, 6) (8) (9)로 분류된다.

군집화 일정표

	결합 군집			처음 나타나는 군집의 단계		
단계	군집 1	군집 2	계수	군집 1	군집 2	다음 단계
1	2	5	.058	0	0	2
2	2	7	.088	1	0	3
3	2	4	.199	2	0	6
4	3	6	.263	0	0	5
5	1	3	.605	0	4	6
6	1	2	1.059	5	3	7
7	1	8	3.559	6	0	8
8	1	9	13.553	7	0	0

소속군집

케이스	5 군집	4 군집	3 군집
1:china	1	1	1
2:cuba	2	2	1
3:Egypt	3	1	1
4:India	2	2	1
5:Israel	2	2	1
6:Netherla	3	1	1
7:Poland	2	2	1
8:U.K	4	3	2
9:U.S	5	4	3

수직 고드름도표

군집의 수	케이스																
	9:U.S		8:U.K		4:India		7:Poland		5:Israel		2:cuba		6:Netherla		3:Egypt		1:china
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X		X		X	X	X	X	X	X	X	X	X	X	X	X	X
4	X		X		X	X	X	X	X	X	X		X	X	X	X	X
5	X		X		X	X	X	X	X	X	X		X	X	X		X
6	X		X		X	X	X	X	X	X	X		X		X		X
7	X		X		X		X	X	X	X	X		X		X		X
8	X		X		X		X		X	X	X		X		X		X

```
* * * * * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * * * * *

 Dendrogram using Single Linkage

                         Rescaled Distance Cluster Combine

    C A S E      0         5        10        15        20        25
  Label     Num  +---------+---------+---------+---------+---------+
  cuba        2   ─┐
  Israel      5   ─┤
  Poland      7   ─┼─┐
  India       4   ─┘ ├──────────┐
  Egypt       3   ─┬─┤          │
  Netherla    6   ─┘ │          ├─────────────────────────────────────┐
  china       1   ───┘          │                                     │
  U.K         8   ──────────────┘                                     │
  U.S         9   ────────────────────────────────────────────────────┘
```

(2) K-평균 군집분석

앞의 예제를 이용하여 상호배반적인 군집분석의 예인 K-평균 군집분석을 실시하였다.

분석-> 분류분석-> K-평균 군집분석
변수 : gnp, trade, defense, gdp
케이스 설명 기준변수 : nation(문자타입)
군집수 : 2 (이부분은 연구자의 의도에 따라서 변경할 수 있다)

옵션 : 분산분석표
방법: 군집방법-가장 가까운 항목(최단연결법), 측도의 간격(제곱유클리디안 거리), 값변환(Z 점수)

해당군집의 내용을 저장하고자 한다면 저장 에서 소속군집 을 설정

그 결과는 다음과 같다.
먼저 초기 군집의 중심은 아래의 표과 같이 주어지며 반복계산에 의한 최종적인 군집의 중심은 gnp는 군집1은 326.31, 군집 1는 2334가 된다.

초기 군집중심

	군집	
	1	2
gnp	58.00	2334.00
trade	349.00	26836.00
defense	3054.00	40641.00
gpd	8.70	12.20

반복계산정보[a]

	군집중심의 변화량	
반복계산	1	2
1	3904.208	.000
2	.000	.000

a. 군집 중심값의 변화가 없거나 작아 수렴이 일어났습니다. 모든 중심에 대한 최대 절대 좌표 변경은 .000입니다. 현재 반복계산은 2입니다. 초기 중심 간의 최소 거리는 46038.288입니다.

최종 군집중심

	군집	
	1	2
gnp	326.21	2334.00
trade	3683.63	26836.00
defense	1041.25	40641.00
gpd	4.64	12.20

또한 각 군집의 변수평균들이 동일하다는 가정을 검정하기 위하여 분산분석을 실시한 결과 아래의 네변인에 대한 군집은 모두 통계적으로 유의한 차이가 있음을 알 수 있다. 이 결과는 단지 비교의 목적에서만 유용하게 사용된다.
즉 이경우에 군집1은 8개의 나라가 군집2는 단지 1개의 나라만 설정된다. 또한 군집2는 U.S가 해당됨을 알 수 있다.

분산분석

	군집		오차		F	유의확률
	평균제곱	자유도	평균제곱	자유도		
gnp	3583298.4	1	112013.790	7	31.990	.001
trade	476473305	1	30264107	7	15.744	.005
defense	1.394E+09	1	2370577.9	7	588.001	.000
gpd	50.837	1	8.086	7	6.287	.041

다른 군집의 여러 케이스 간 차이를 최대화하기 위해 군집을 선택했으므로 F 검정은 기술 통계를 목적으로만 사용되어야 합니다. 이 경우 관측유의수준은 수정되지 않으므로 군집 평균이 동일하다는 가설을 검정하는 것으로 해석될 수 없습니다.

각 군집의 케이스 수

군집	1	8.000
	2	1.000
유효		9.000
결측		.000