

분산분석

분산분석(ANOVA: ANALYSIS OF VARIANCE)

◆ 두 개 이상의 모집단의 차이를 검정

- 예: 회사에서 세 종류의 기계를 설치하여 동일한 제품을 생산하는 경우, 각 기계의 생산량을 조사하여 평균 생산량을 비교
- 독립변수: 다른 변수에 의해 영향을 주는 변수
- 종속변수: 다른 변수에 의해 영향을 받는 변수
- 요인(Factor): 독립변수
- 예에서의 요인: 기계의 종류 (I, II, III)
- 요인수준(Factor level, treatment): 요인내에서 영향을 미치는 형태 (기계 I, 기계 II, 기계III)
- 예에서의 종속변수: 생산량
- 일원분산분석(One factor ANOVA): 요인이 하나인 경우

분산 분석의 가정 및 종류

◆ 분산분석의 가정

- 각 요인수준에 대응하는 모집단은 동일한 분산을 가진다.
- 각 요인수준에 대응하는 모집단은 정규분포이다.
- 각 요인수준에 대한 관찰치들은 임의로 얻어지는 것이며 독립적이다.

◆ 일원분산분석(One factor ANOVA): 요인이 하나인 경우

◆ 이원분산분석 (Two factor ANOVA) : 요인이 두 개인 경우

- 반복이 없는 경우
- 반복이 있는 경우

일원 분산분석 (ONE FACTOR ANOVA)

	기계 I	기계 II	기계 III
생산량	25	21	22
	20	20	20
	25	16	21
	26	15	

일원 분산분석 (ONE FACTOR ANOVA)

표본 (i)	처리 (j)			총계
	1	2	3	
1	$Y_{11} = 25$	$Y_{12} = 21$	$Y_{13} = 22$	
2	$Y_{21} = 20$	$Y_{22} = 20$	$Y_{23} = 20$	
3	$Y_{31} = 25$	$Y_{32} = 16$	$Y_{33} = 21$	
4	$Y_{41} = 26$	$Y_{42} = 15$		
합계	$Y_1 = 96$	$Y_2 = 72$	$Y_3 = 63$	$Y = 231$
평균	$\bar{Y}_1 = 24$	$\bar{Y}_2 = 18$	$\bar{Y}_3 = 21$	$\bar{Y} = 21$
표본갯수	$n_1 = 4$	$n_2 = 4$	$n_3 = 3$	$n = 11$
요인수준 효과	$\alpha_1 = 3$	$\alpha_2 = -3$	$\alpha_3 = 0$	

일원 분산분석 (ONE FACTOR ANOVA)

◆ 변동의 분해

$$\begin{bmatrix} 25 & 21 & 22 \\ 20 & 20 & 20 \\ 25 & 16 & 21 \\ 26 & 15 & \end{bmatrix} = \begin{bmatrix} 21 & 21 & 21 \\ 21 & 21 & 21 \\ 21 & 21 & 21 \\ 21 & 21 & \end{bmatrix} + \begin{bmatrix} 3 & -3 & 0 \\ 3 & -3 & 0 \\ 3 & -3 & 0 \\ 3 & -3 & \end{bmatrix} + \begin{bmatrix} 1 & 3 & 1 \\ -4 & 2 & -1 \\ 1 & -2 & 0 \\ 2 & -3 & \end{bmatrix}$$

(관찰치)

(전체평균) (요인수준효과)

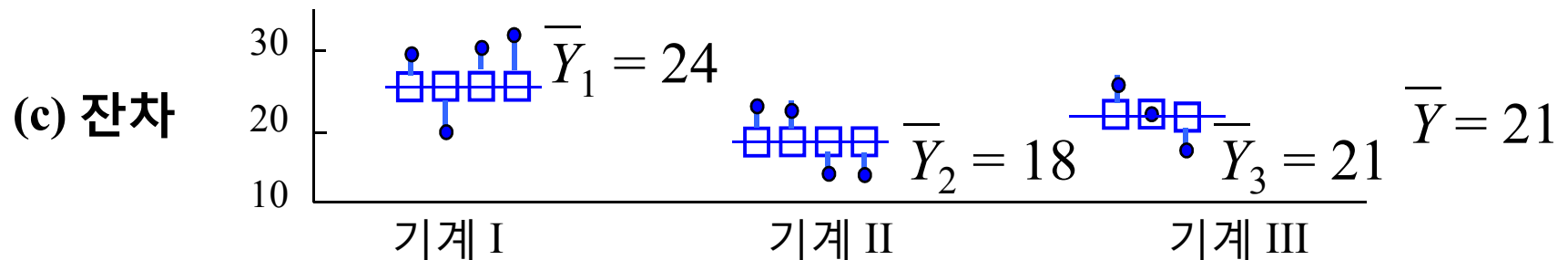
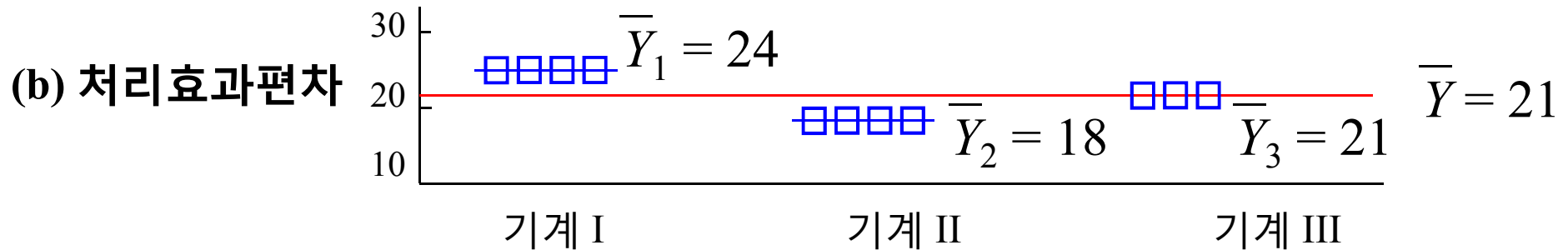
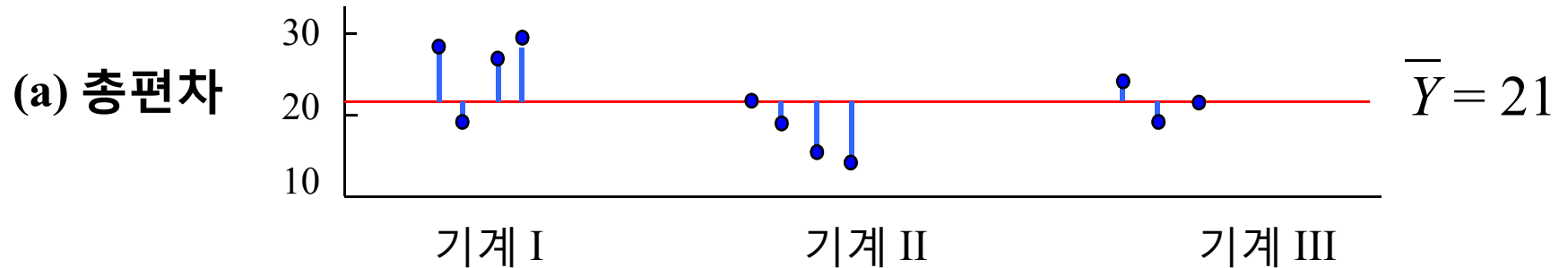
(잔차)

$$\begin{bmatrix} 4 & 0 & 1 \\ -1 & -1 & -1 \\ 4 & -5 & 0 \\ 5 & -6 & \end{bmatrix} = \begin{bmatrix} 3 & -3 & 0 \\ 3 & -3 & 0 \\ 3 & -3 & 0 \\ 3 & -3 & \end{bmatrix} + \begin{bmatrix} 1 & 3 & 1 \\ -4 & 2 & -1 \\ 1 & -2 & 0 \\ 2 & -3 & \end{bmatrix}$$

(총편차)

(요인수준효과) (잔차)

일원 분산분석 (ONE FACTOR ANOVA)



분산분석의 용어

◆ 총변동 (SST: Sum of Squares Total)

- 각 관찰치와 전체 표본 평균의 편차 제곱의 합
- $\sum (Y_{ij} - \bar{Y})^2 = (25-21)^2 + (20-21)^2 + \dots + (21-21)^2 = 122$

◆ 그룹간 변동 (SSB: Sum of Squares Between groups)

- (각 그룹의 평균과 전체 표본 평균의 편차 제곱) * 그룹의 표본크기 의 합
- $\sum n_j (\bar{Y}_j - \bar{Y})^2 = 4 (24-21)^2 + 4 (18-21)^2 + 3 (21-21)^2 = 72$

◆ 그룹내 변동 (SSW: Sum of Squares Within groups)

- 그룹내 관찰치와 그룹의 평균간의 편차 제곱합
- $\sum \sum n_j (Y_{ij} - \bar{Y}_j)^2 = \{(25-24)^2 + \dots + (26-24)^2\} + \{(21-18)^2 + \dots + (15-18)^2\} + \{(22-21)^2 + \dots + (21-21)^2\} = 50$

◆ SST = SSB + SSW

◆ 그룹간 평균제곱 (MSB: Mean Squares Between groups): MSB = SSB/(g-1)

◆ 그룹내 평균제곱 (MSW: Mean Squares Within groups) : MSW = SSW/(n-g)

분산분석의 가설검정

- ◆ H_0 : 모든 그룹의 평균은 같다. (요인수준에 따른 차이가 없다)
- ◆ H_1 : 모든 그룹의 평균이 다 같은 것은 아니다. (평균이 서로 다른 그룹이 존재한다. 요인수준에 따른 차이가 있다)

- ◆ IF p-value > 유의수준, Then H_0 채택
- ◆ IF p-value < 유의수준, Then H_0 기각, H_1 채택 => 서로 다른 그룹을 찾아냄 (Post hoc analysis, 사후분석 시행)

사후분석

◆ Fisher's Least Significant Difference

- 두 수준별 평균비교 검정에 사용한다. LSD를 구하고 평균의 차이가 그보다 크면 귀무가설을 기각한다.

◆ Tukey

- 가장 보수적인 방법으로 **자연과학**에서 많이 사용

◆ Student-Newman-Keuls procedure

- Tukey 와 결과 동일

◆ Duncan Multiple range test

- Tukey와 유사, 수준별 표본 평균의 크기 순으로 나열하여 차이가 큰 것을 비교해 가면서 유의 수준을 $1-(1-\alpha)r$ 으로 조정해 가면서 검정. R은 검정단계 순서. 귀무가설을 기각할 가능성이 높음

◆ **Scheffe's S Method**

- **사회과학**에 많이 사용

반복측정이 없는 분산분석

◆ 생산 실적표

작업자 \ 기계	기계 I	기계 II	기계III	합	평균
1년	25	20	21	66	22
4년	28	22	19	69	23
8년	22	18	23	63	21
합	75	60	63	198	
평균	25	20	21		22

반복측정이 없는 이원분산분석표

원천	제공합	자유도	평균제곱	F
요인 1(A)	$SSA = c \sum_{i=1}^g (\bar{Y}_i - \bar{Y})^2$	g-1	MSB =SSA/(g-1)	MSA/MSE
요인 2(B)	$SSB = g \sum_{j=1}^c (\bar{Y}_j - \bar{Y})^2$	c-1	MSW =SSB/(c-1)	MSB/MSE
잔차	$SSW = \sum_{i=1}^g \sum_{i=1}^c (Y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y})^2$	(g-1)(c-1)	$MSE = \frac{SSW}{(g-1)(c-1)}$	
합계	$SST = \sum_{i=1}^g \sum_{i=1}^c (Y_{ij} - \bar{Y})^2$	gc-1		

◆ $SST = (25-22)^2 + (28-22)^2 + \dots + (23-22)^2 = 76$

◆ $SSA = 3\{(25-22)^2 + (20-22)^2 + (21-22)^2\} = 42$

◆ $SSB = 3\{(22-22)^2 + (23-22)^2 + (21-22)^2\} = 6$

◆ $SSW = (25-25-22+22)^2 + (28-25-23+22)^2 + \dots + (23-21-21+22)^2 = 28$

• $SST = SSA + SSB + SSW$

반복이 있는 이원분산분석 모형

◆ 화학공장의 수율자료

온도 \ 압력	압력		
	200	250	300
저온	98	108	104
	89	99	111
	86	114	100
고온	99	115	106
	102	109	99
	102	121	92

상호작용효과 및 가설

- ◆ 하나의 요인이 다른 요인의 변화에 영향을 미침
- ◆ 요인의 변화에 따른 기대반응치의 변화를 분석함
 - 프로파일 작성
 - 상호교차점이 있거나 평행에서 많이 벗어나는 경우 상호작용을 있다고 추측
 - 상호작용이 없을 경우 상호작용항을 제거
- ◆ 상호작용
 - H_0 : 모든 상호작용 = 0 이다. (상호작용이 없다)
 - H_1 : 모든 상호작용 = 0 인 것은 아니다. (상호작용이 있다)

실습1- 일원분산분석

- ◆ 다음 세 종류의 기계에서 생산되는 생산량의 차이가 있는지 여부를 유의수준 0.05에서 검정하시오.

	기계A	기계B	기계X
생산량	25	21	22
	20	20	20
	25	16	20
	26	15	21

- ◆ 실습파일 (anova1.sav)을 이용

Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

11 :

	기계	생산량
1	1.00	25.00
2	1.00	20.00
3	1.00	25.00
4	1.00	26.00
5	2.00	21.00
6	2.00	20.00
7	2.00	16.00
8	2.00	15.00
9	3.00	22.00
10	3.00	20.00
11	3.00	21.00
12		
13		
14		
15		

Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples

Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...

Data View / Variable View /
One-Way ANOVA SPSS Processor is ready

One-Way ANOVA

Dependent List:
 생산량

Factor:
 기계

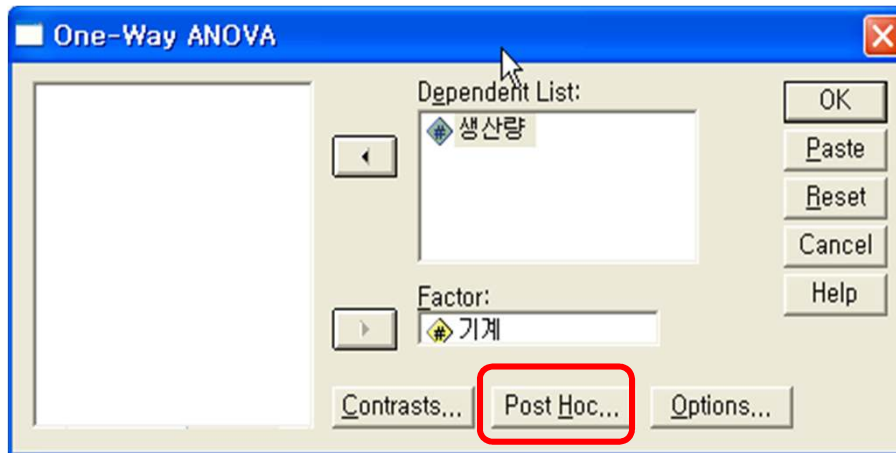
OK
Paste
Reset
Cancel
Help

Contrasts... Post Hoc... Options...

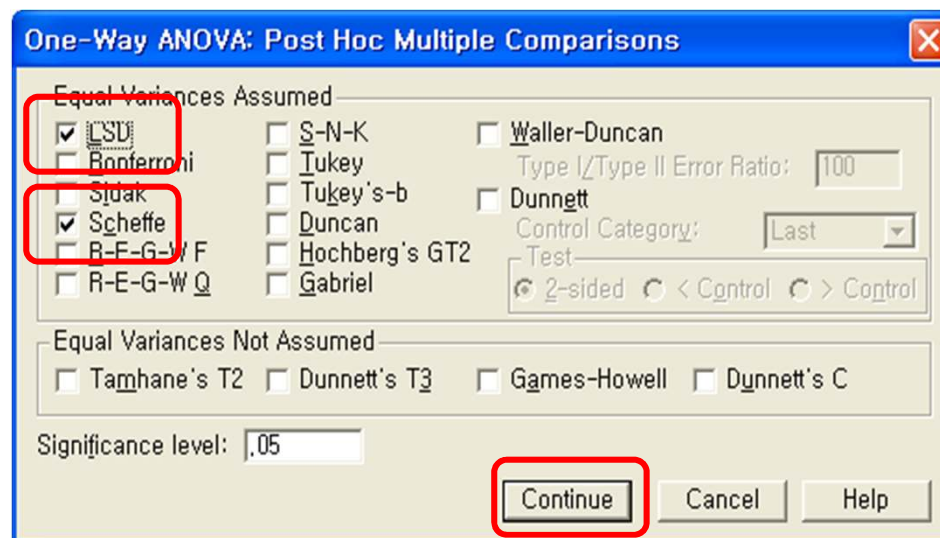
ANOVA

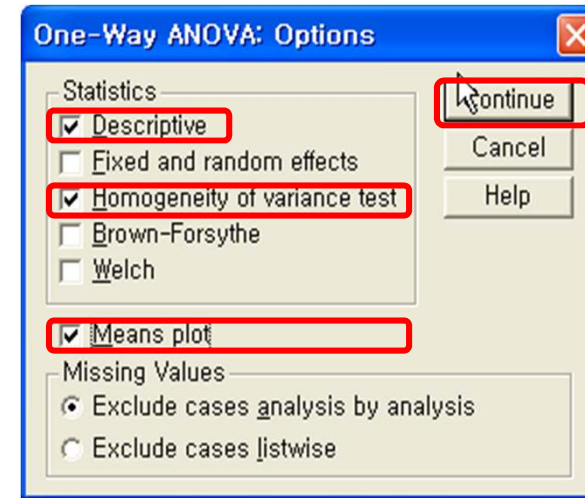
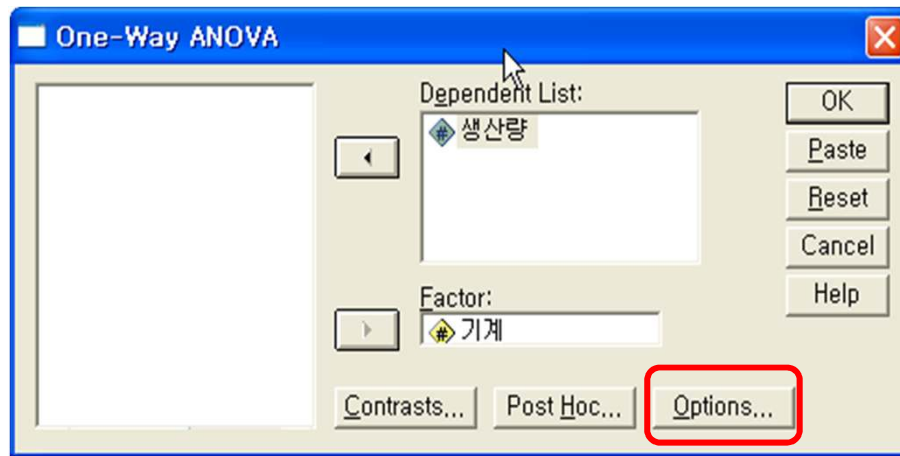
생산량

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	72.000	2	36.000	5.760	.028
Within Groups	50.000	8	6.250		
Total	122.000	10			



H_0 : 평균이 동일
 H_1 : 평균이 다르다





Test of Homogeneity of Variances

생산량

Levene Statistic	df1	df2	Sig.
3.115	2	8	.100

H_0 : 모분산이 동일하다
 H_1 : 모분산이 동일하지 않다

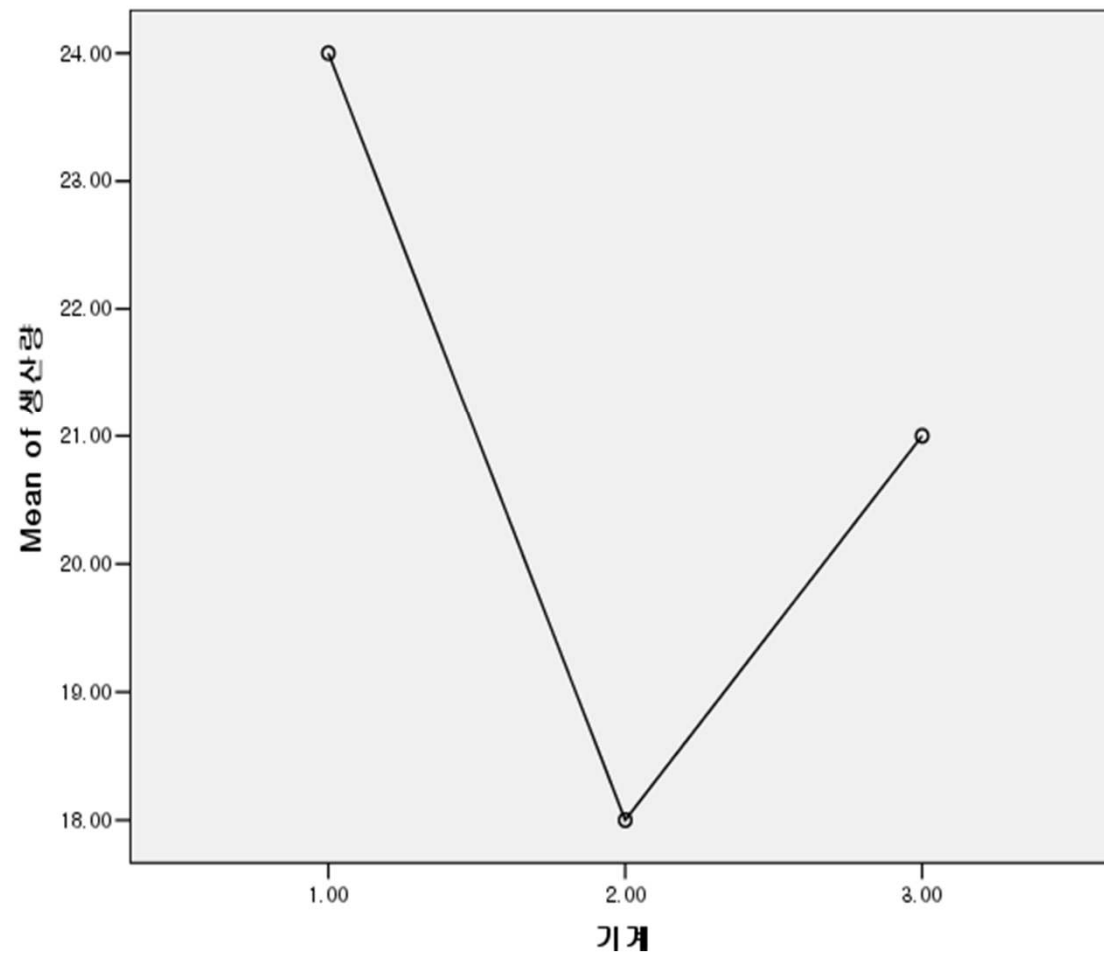
Multiple Comparisons

Dependent Variable: 생산량

	(I) 기계	(J) 기계	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	1.00	2.00	6.00000*	1.76777	.028	.7209	11.2791
		3.00	3.00000	1.90941	.341	-2.7020	8.7020
	2.00	1.00	-6.00000*	1.76777	.028	-11.2791	-.7209
		3.00	-3.00000	1.90941	.341	-8.7020	2.7020
	3.00	1.00	-3.00000	1.90941	.341	-8.7020	2.7020
		2.00	3.00000	1.90941	.341	-2.7020	8.7020
LSD	1.00	2.00	6.00000*	1.76777	.009	1.9235	10.0765
		3.00	3.00000	1.90941	.155	-1.4031	7.4031
	2.00	1.00	-6.00000*	1.76777	.009	-10.0765	-1.9235
		3.00	-3.00000	1.90941	.155	-7.4031	1.4031
	3.00	1.00	-3.00000	1.90941	.155	-7.4031	1.4031
		2.00	3.00000	1.90941	.155	-1.4031	7.4031

*. The mean difference is significant at the .05 level.

Means Plots



실습2 - 반복이 없는 이원분산분석

- ◆ 기계종류와 작업자의 경력수준에 따라 생산량의 차이가 있는지 유의수준 0.05에서 검정하시오.

작업자 \ 기계	기계 I	기계 II	기계III
1년	25	20	21
4년	28	22	19
8년	22	18	23

- ◆ 실습파일 (anova2.sav)을 이용

Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

10 : 생산량

	기계	작업자	생산량
1	1.00	1.00	25.00
2	1.00	2.00	28.00
3	1.00	3.00	22.00
4	2.00	1.00	20.00
5	2.00	2.00	22.00
6	2.00	3.00	18.00
7	3.00	1.00	21.00
8	3.00	2.00	19.00
9	3.00	3.00	23.00
10			
11			
12			
13			
14			
15			

General Linear Model > Univariate...

SPSS Processor

Univariate

Dependent Variable: 생산량

Fixed Factor(s): 기계, 작업자

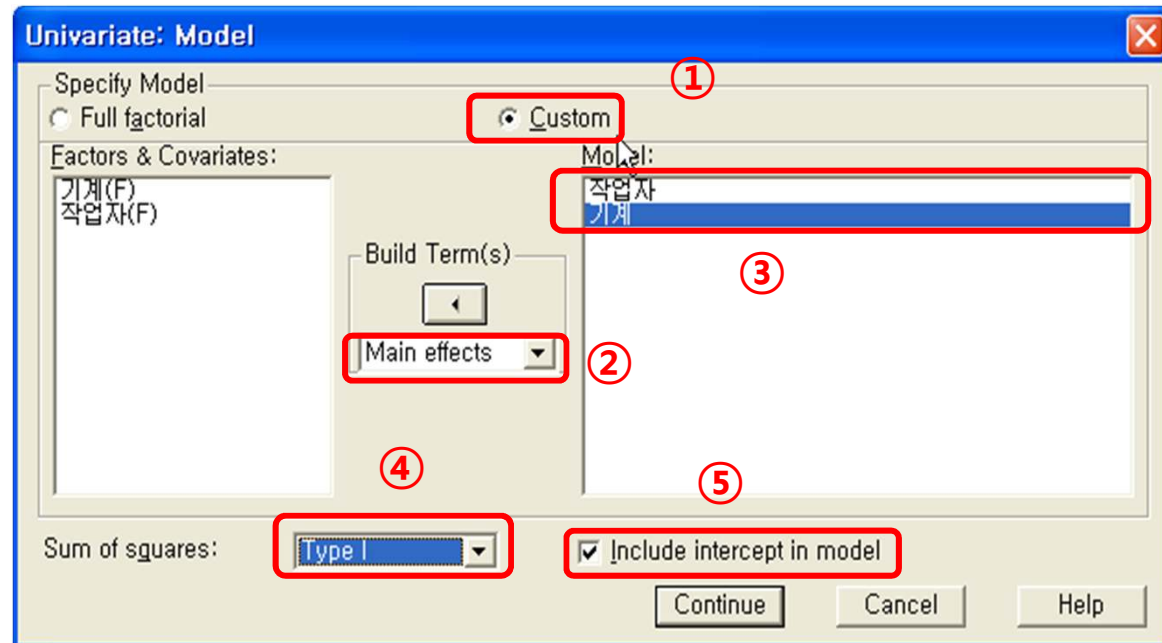
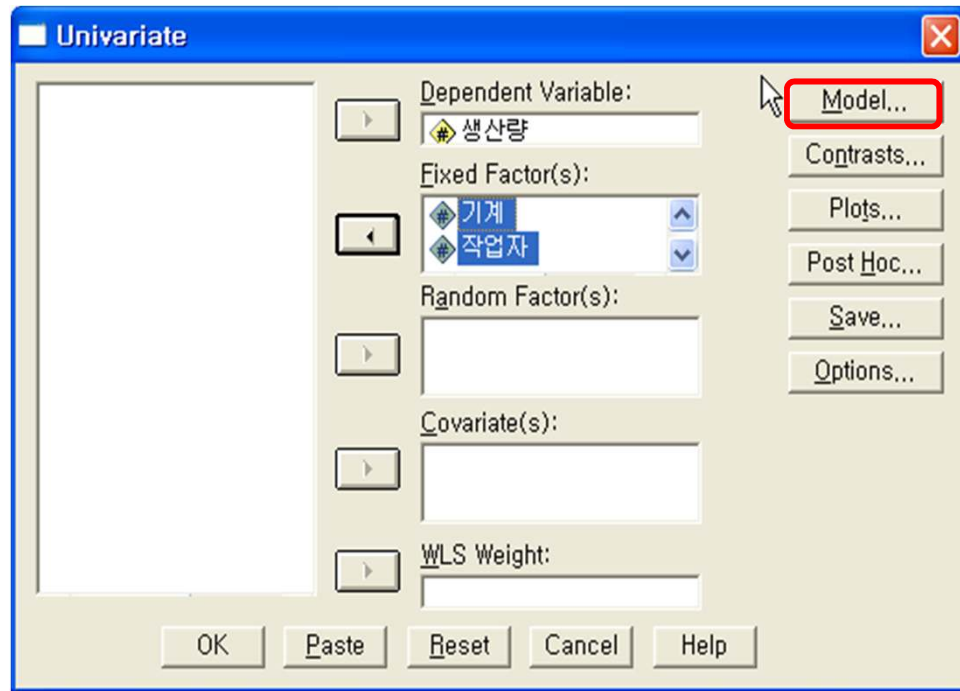
Random Factor(s):

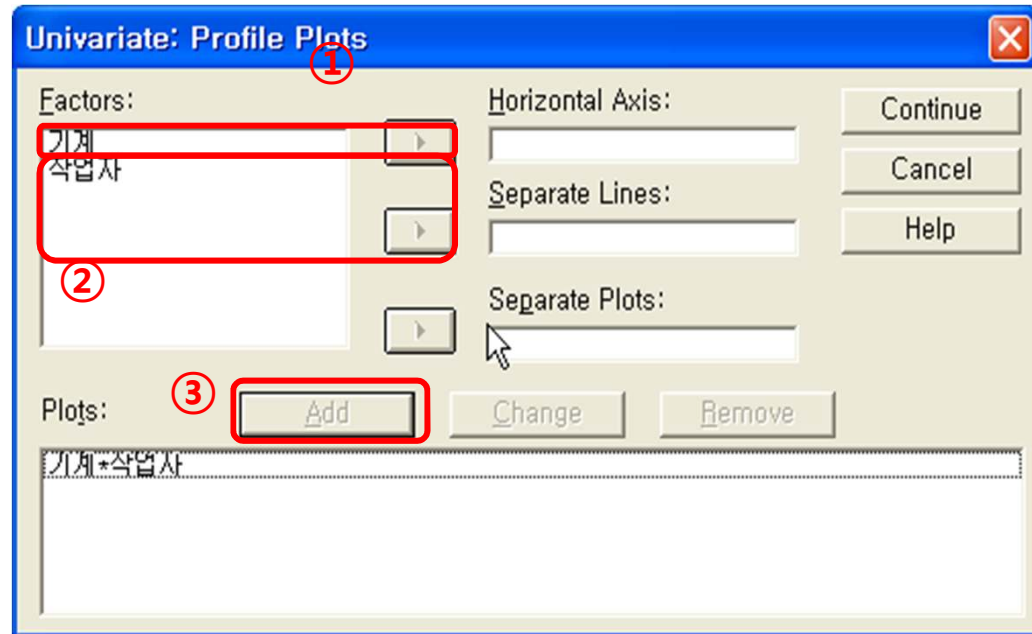
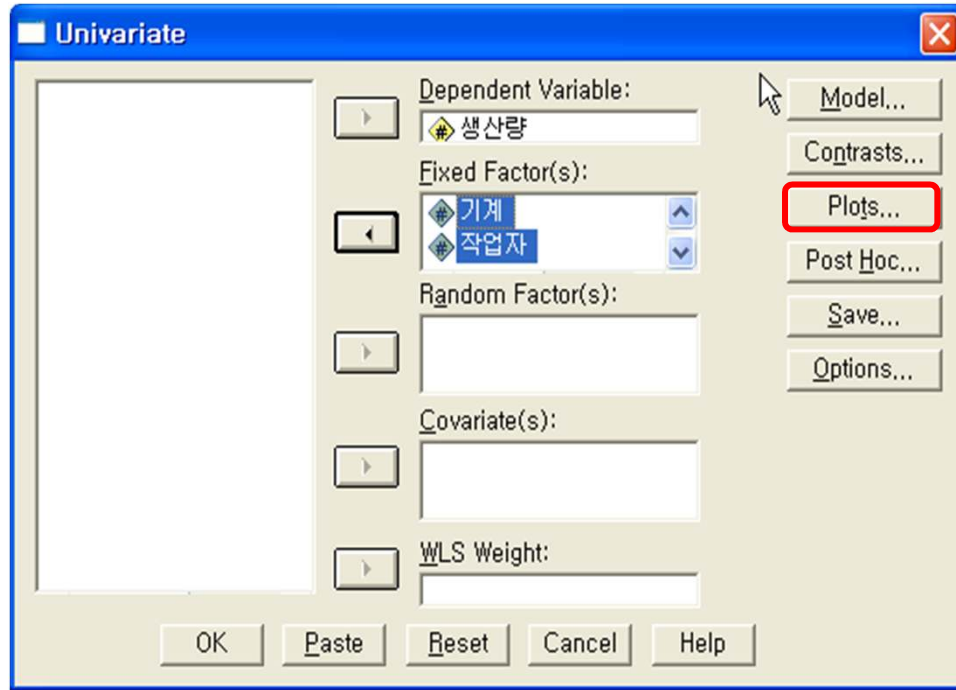
Covariate(s):

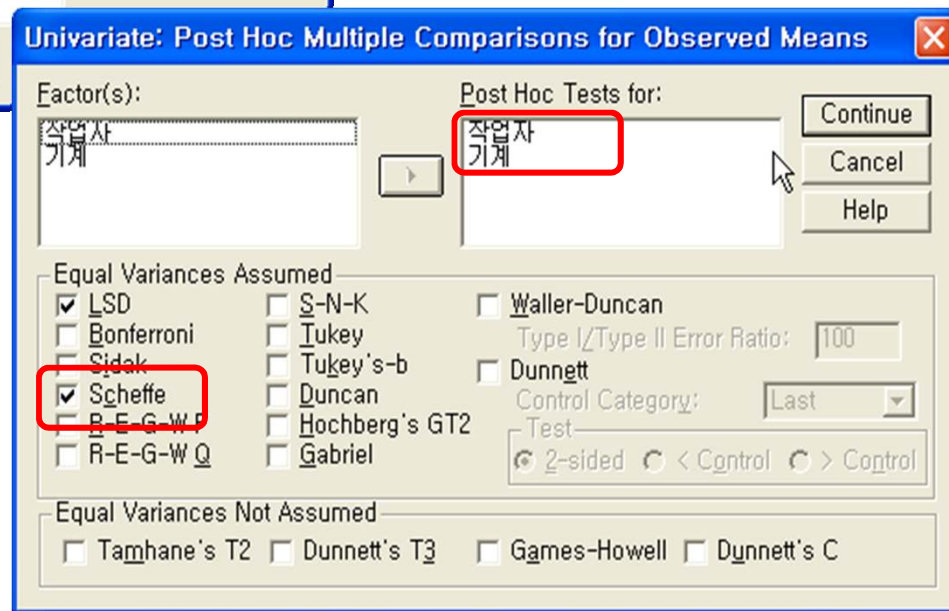
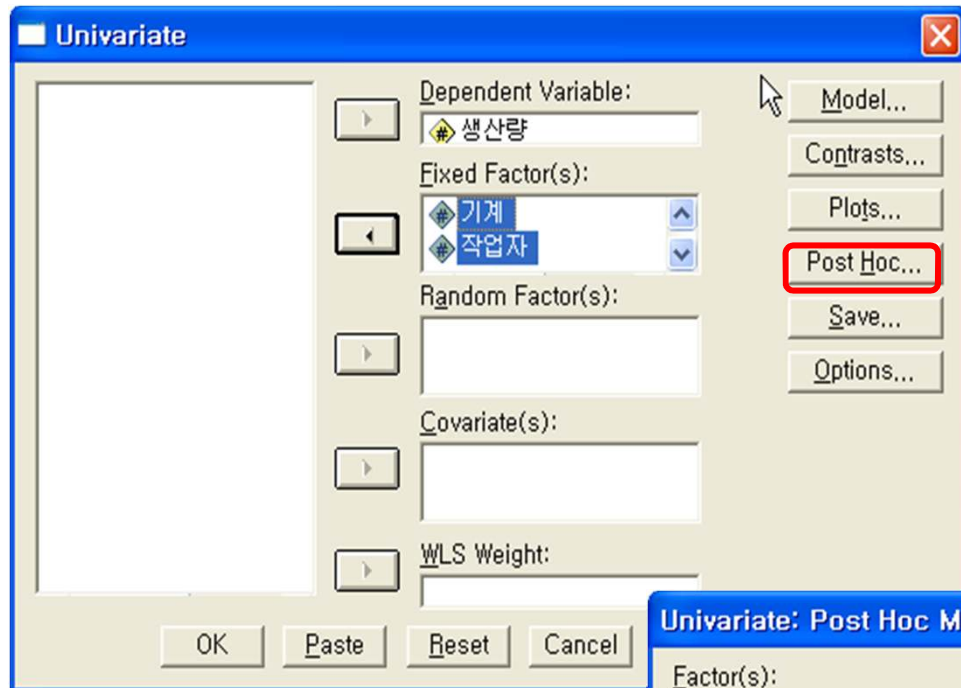
WLS Weight:

Model...
Contrasts...
Plots...
Post Hoc...
Save...
Options...

OK Paste Reset Cancel Help







Tests of Between-Subjects Effects

Dependent Variable: 생산량

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	48.000 ^a	4	12.000	1.714	.307
Intercept	4356.000	1	4356.000	622.286	.000
작업자	6.000	2	3.000	.429	.678
기계	42.000	2	21.000	3.000	.160
Error	28.000	4	7.000		
Total	4432.000	9			
Corrected Total	76.000	8			

a. R Squared = .632 (Adjusted R Squared = .263)

H_0 : 작업자별
평균생산량이
동일

H_0 기계별 평균
생산량이 동일

Multiple Comparisons

Dependent Variable: 생산량

	(I) 작업자	(J) 작업자	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	1.00	2.00	-1.0000	2.16025	.901	-9.0507	7.0507
		3.00	1.0000	2.16025	.901	-7.0507	9.0507
	2.00	1.00	1.0000	2.16025	.901	-7.0507	9.0507
		3.00	2.0000	2.16025	.678	-6.0507	10.0507
	3.00	1.00	-1.0000	2.16025	.901	-9.0507	7.0507
		2.00	-2.0000	2.16025	.678	-10.0507	6.0507
LSD	1.00	2.00	-1.0000	2.16025	.667	-6.9978	4.9978
		3.00	1.0000	2.16025	.667	-4.9978	6.9978
	2.00	1.00	1.0000	2.16025	.667	-4.9978	6.9978
		3.00	2.0000	2.16025	.407	-3.9978	7.9978
	3.00	1.00	-1.0000	2.16025	.667	-6.9978	4.9978
		2.00	-2.0000	2.16025	.407	-7.9978	3.9978

Based on observed means.

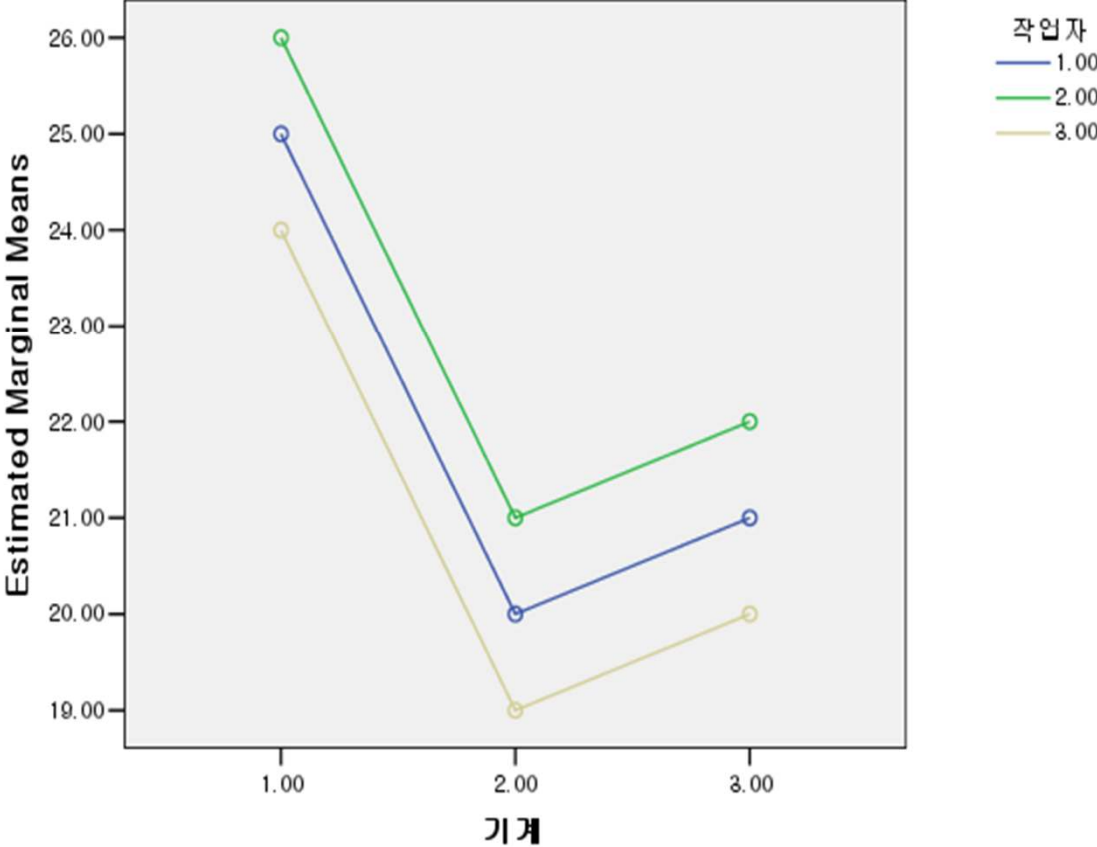
Multiple Comparisons

Dependent Variable: 생산량

	(I) 기계	(J) 기계	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	1.00	2.00	5.0000	2.16025	.183	-3.0507	13.0507
		3.00	4.0000	2.16025	.290	-4.0507	12.0507
	2.00	1.00	-5.0000	2.16025	.183	-13.0507	3.0507
		3.00	-1.0000	2.16025	.901	-9.0507	7.0507
	3.00	1.00	-4.0000	2.16025	.290	-12.0507	4.0507
		2.00	1.0000	2.16025	.901	-7.0507	9.0507
LSD	1.00	2.00	5.0000	2.16025	.082	-.9978	10.9978
		3.00	4.0000	2.16025	.138	-1.9978	9.9978
	2.00	1.00	-5.0000	2.16025	.082	-10.9978	.9978
		3.00	-1.0000	2.16025	.667	-6.9978	4.9978
	3.00	1.00	-4.0000	2.16025	.138	-9.9978	1.9978
		2.00	1.0000	2.16025	.667	-4.9978	6.9978

Based on observed means.

Estimated Marginal Means of 생산량



실습 3 - 반복이 있는 이원분산분석

- ◆ 화학공장의 수율이 다음과 같을 때 온도와 압력에 따른 수율의 차이가 있는지 유의수준 0.05에서 검정하시오.

온도 \ 압력	압력		
	200	250	300
저온	98	108	104
	89	99	111
	86	114	100
고온	99	115	106
	102	109	99
	102	121	92

- ◆ 실습파일 (anova3.sav)을 이용

Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

10 : 생산량

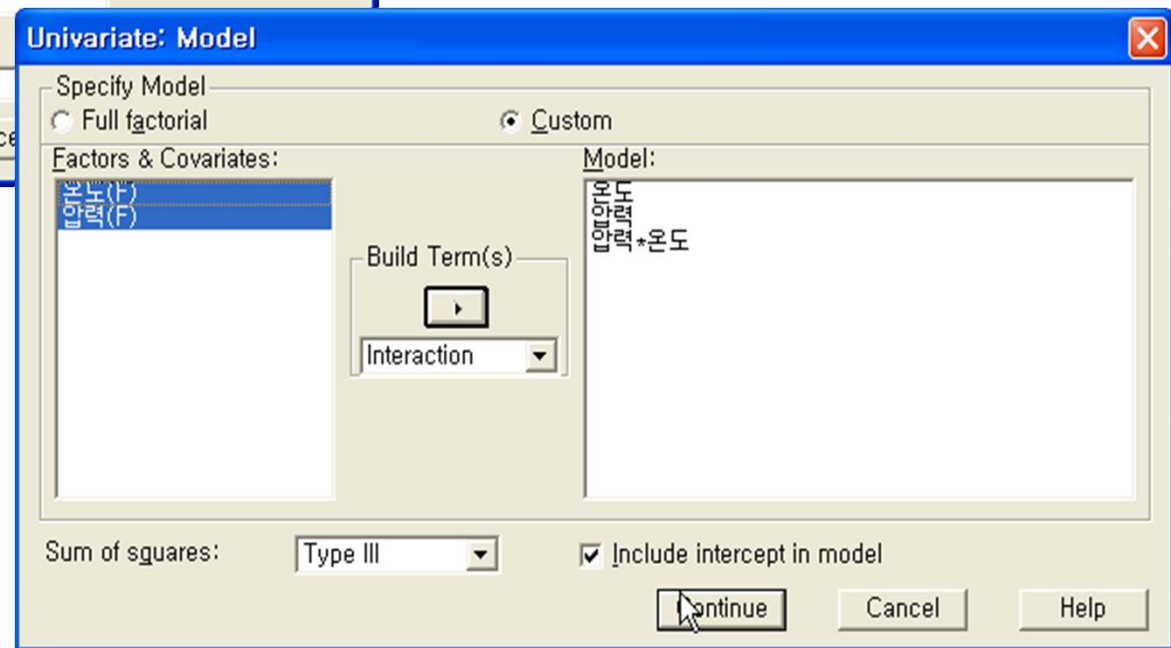
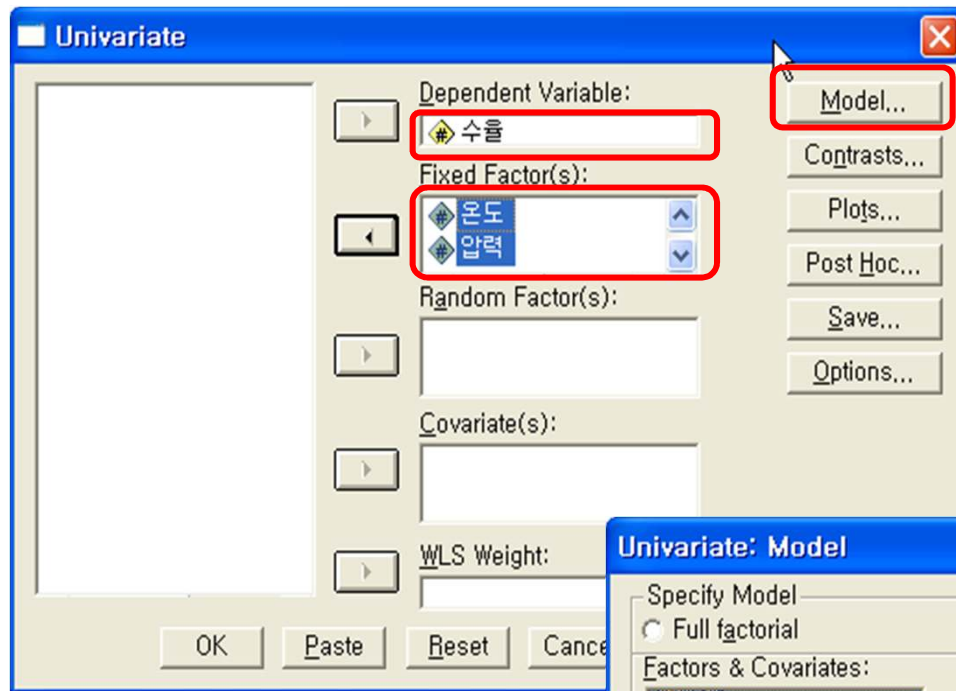
	기계	작업자	생산량
1	1.00	1.00	25.00
2	1.00	2.00	28.00
3	1.00	3.00	22.00
4	2.00	1.00	20.00
5	2.00	2.00	22.00
6	2.00	3.00	18.00
7	3.00	1.00	21.00
8	3.00	2.00	19.00
9	3.00	3.00	23.00
10			
11			
12			
13			
14			
15			

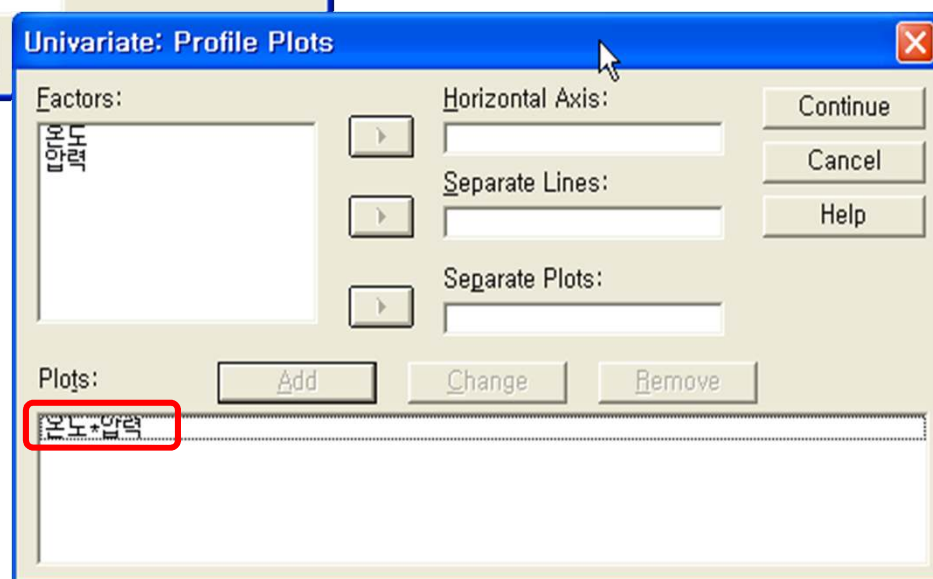
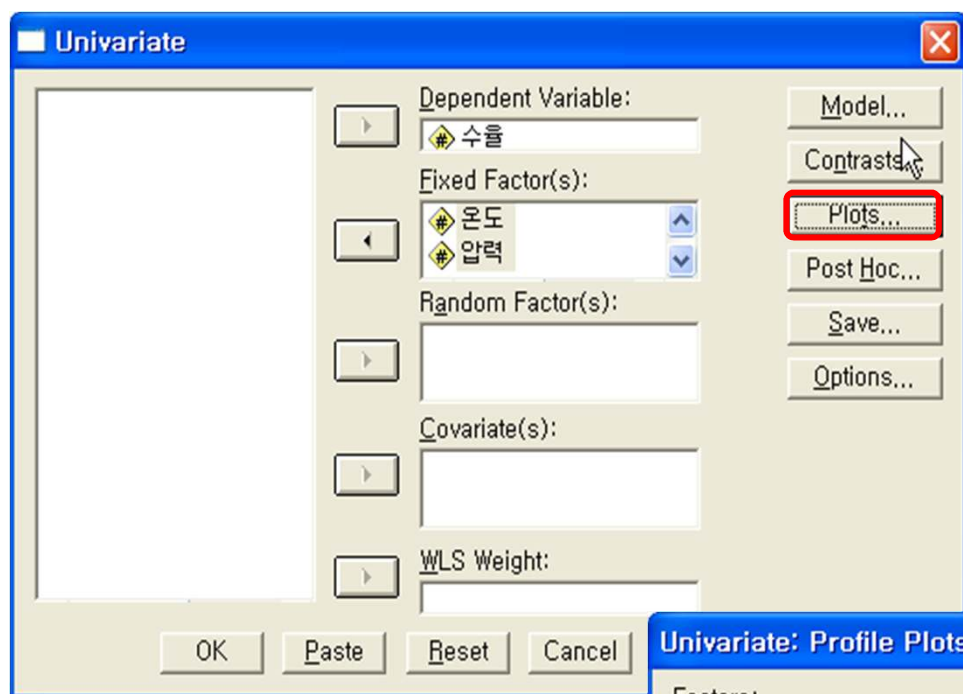
Reports
 Descriptive Statistics
 Tables
 Compare Means
General Linear Model
 Mixed Models
 Correlate
 Regression
 Loglinear
 Classify
 Data Reduction
 Scale
 Nonparametric Tests
 Time Series
 Survival
 Multiple Response
 Missing Value Analysis...
 Complex Samples

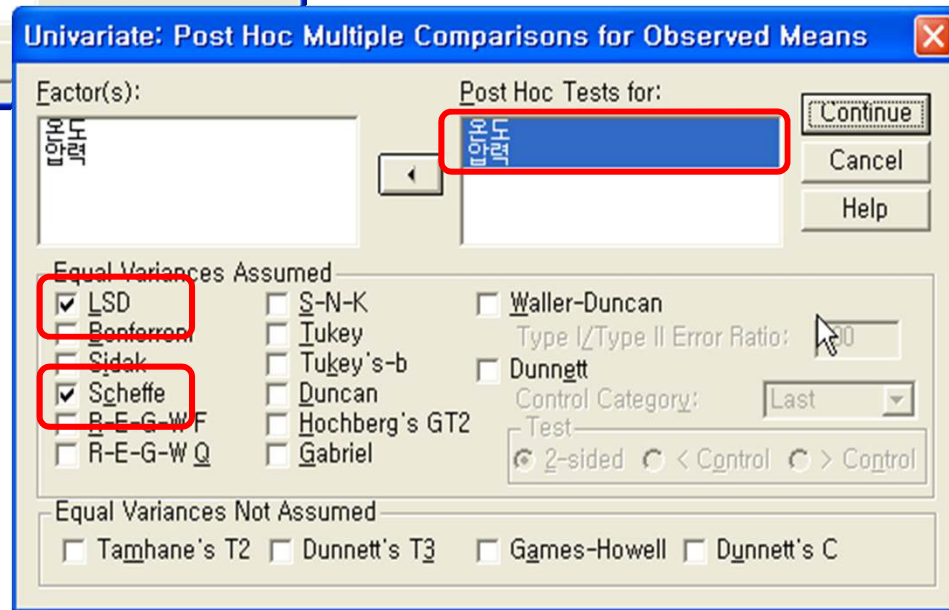
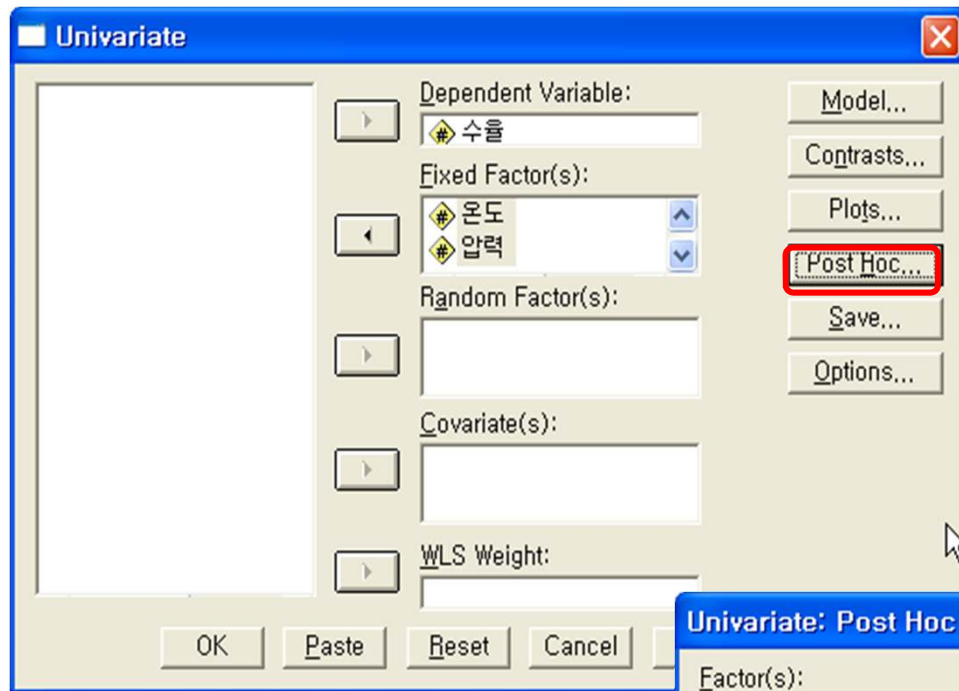
Univariate...
 Multivariate...
 Repeated Measures...
 Variance Components...

Data View / Variable View /

General Factorial SPSS Processor is ready







Tests of Between-Subjects Effects

Dependent Variable: 수율

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	984.000 ^a	5	196.800	5.492	.007
Intercept	190962.000	1	190962.000	5329.172	.000
온도	72.000	1	72.000	2.009	.182
압력	684.000	2	342.000	9.544	.003
온도 * 압력	228.000	2	114.000	3.181	.078
Error	430.000	12	35.833		
Total	192376.000	18			
Corrected Total	1414.000	17			

a. R Squared = .696 (Adjusted R Squared = .569)

H_0 : 온도별 평균수율이 동일

H_1 : 압력별 평균수율이 동일하지 않다.

H_0 : 온도와 압력간 상호작용은 없다

Multiple Comparisons

Dependent Variable: 수율

	(I) 압력	(J) 압력	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	1.00	2.00	-15.0000*	3.45607	.003	-24.6341	-5.3659
		3.00	-6.0000	3.45607	.261	-15.6341	3.6341
	2.00	1.00	15.0000*	3.45607	.003	5.3659	24.6341
		3.00	9.0000	3.45607	.068	-.6341	18.6341
	3.00	1.00	6.0000	3.45607	.261	-3.6341	15.6341
		2.00	-9.0000	3.45607	.068	-18.6341	.6341
LSD	1.00	2.00	-15.0000*	3.45607	.001	-22.5301	-7.4699
		3.00	-6.0000	3.45607	.108	-13.5301	1.5301
	2.00	1.00	15.0000*	3.45607	.001	7.4699	22.5301
		3.00	9.0000*	3.45607	.023	1.4699	16.5301
	3.00	1.00	6.0000	3.45607	.108	-1.5301	13.5301
		2.00	-9.0000*	3.45607	.023	-16.5301	-1.4699

Based on observed means.

*. The mean difference is significant at the .05 level.

Estimated Marginal Means of 수출

