제7장 분산분석

평균의 비교: 독립표본 t 검정과 분산분석

- 독립표본 t 검정 (모집단 2개)
  - $H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$
- 분산분석 (모집단 2개 이상)
- 모집단이 3개인 경우의 가설
  - $H_0 : \mu_1 = \mu_2 = \mu_3$ $H_1 : H_0$ 가 아니다.
  - H_0 가 아니다 $\Leftrightarrow \mu_1 \neq \mu_2 \neq \mu_3$?? **No!!**
  - 따라서 H_1 은 다음과 같은 의미
    - $\mu_1 \neq \mu_2 \neq \mu_3$ 또는
    - $\mu_1 = \mu_2 \neq \mu_3$ 또는
    - $\mu_1 \neq \mu_2 = \mu_3$ 또는
    - $\mu_1 \neq \mu_3 = \mu_2$

다중비교와 1종 오류

- **다중비교와 1종 오류**
  - **만일 ANOVA 가 아닌 여러 번 t 검정을 하면 안되나?**
    - **1종 오류가능성이 증대됨**
    - **예) 세 집단을 비교하기 위해서는 세 번의 독립표본 t 검정을 수행하여야 함.**
    - **각 t 검정에서 유의수준을 0.05로 설정하였다면, 세 번 모두 귀무가설이 맞는데 귀무가설을 기각하지 않은 옳은 결정을 할 확률은**
      - **$0.95 \times 0.95 \times 0.95 = 0.86$, 1종 오류는 0.14**
    - **연습문제 m개의 집단을 두 개씩 t 검정을 수행한다면 1종오류는?**
      - **$1-(1-\alpha)^{m(m-1)/2}$**

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

일원 분산분석 (One factor ANOVA)

- **점포별 매출액 (백만원)**

	점포 I	점포 II	점포 III
매출액	25 20 25 26	21 20 16 15	22 20 21

- **종속변수 : 매출액**
- **요인 : 점포**
- **요인수준 : 점포 I, 점포 II, 점포 III**

분산분석의 원리(1/2)

- **총변동 (SST: Sum of Squares Total)**
  - **각 관찰치와 전체 표본 평균의 편차 제곱의 합**
  - $\Sigma(Y_{ij} - Y)^2 = (25-21)^2 + (20-21)^2 + \ldots + (21-21)^2 = 122$
- **그룹간 변동 (SSB: Sum of Squares Between groups)**
  - **(각 그룹의 평균과 전체 표본 평균의 편차 제곱)*그룹의 표본크기 의 합**
  - $\Sigma\, n_j\, (Y_j - Y)^2 = 4\,(24-21)^2 + 4\,(18-21)^2 + 3\,(21-21)^2 = 72$
- **그룹내 변동 (SSW: Sum of Squares Within groups)**
  - **그룹내 관찰치와 그룹의 평균간의 편차 제곱합**
  - $\Sigma\Sigma\, n_j\, (Y_{ij} - Y_j)^2 = \{(25-24)^2 + \ldots + (26-24)^2\} + \{(21-18)^2 + \ldots + (15-18)^2\} + \{(22-21)^2 + \ldots + (21-21)^2\} = 50$
- **SST = SSB + SSW**

F-test

- **F값과 p-value**

엑셀 활용

m개의 범주에서 n개의 표본이 추출되었다면, 관련된 F 분포의 자유도는 $m-1$과 $n-m$

❑ $Pr(F > x)$를 구할 때, =FDIST ($x, m-1, n-m$)

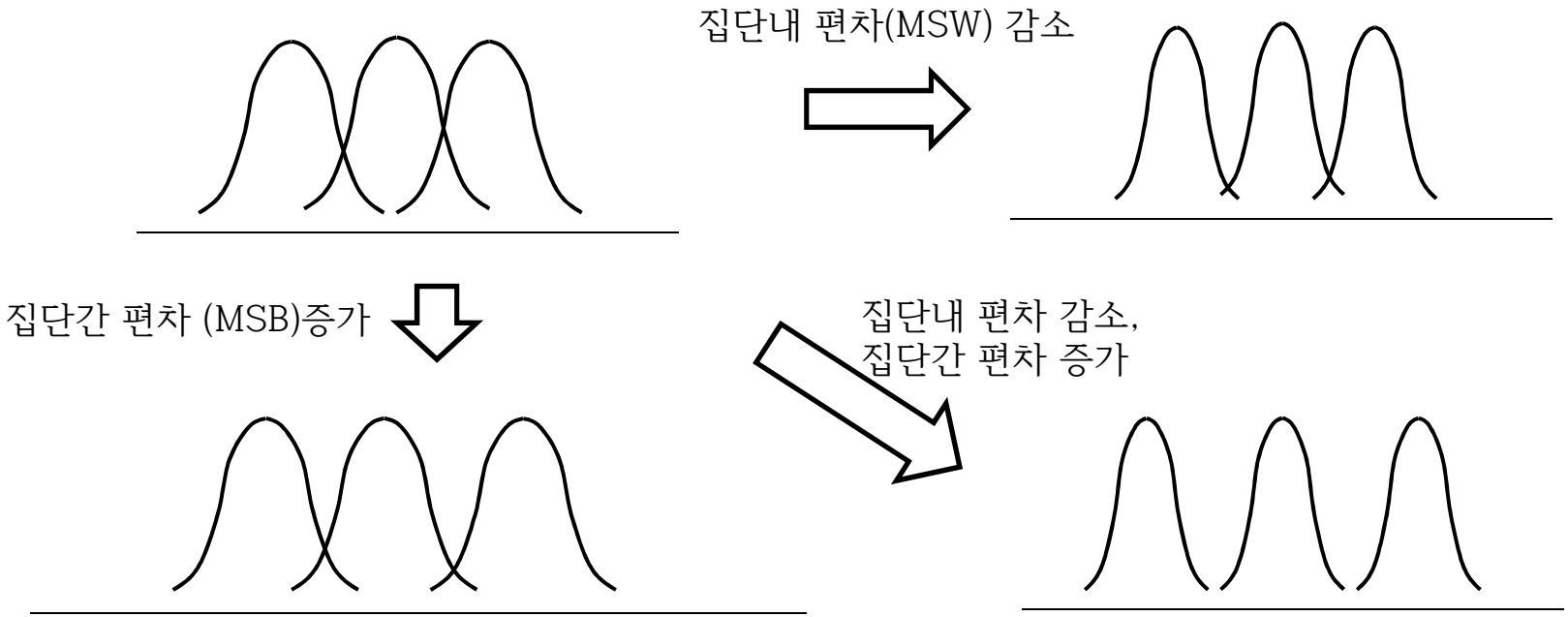
이 함수가 $Pr(F < x)$를 구하는 것이 아님에 유의

❑ $Pr(F > x) = p$ 인 x를 구할 때, =FINV ($p, m-1, n-m$)

이 함수는 역함수로서 p를 구하는 것이 아니고 x를 구하는 것임에 유의

분산분석의 원리(2/2)

- **집단간 편차(MSB: Mean Squares Between groups): MSB = SSB/(g-1)**
- **집단내 편차(MSW: Mean Squares Within groups) : MSW = SSW/(n-g)**
- **F 값 =MSB/MSW**
  - **의미 : F값은 집단내의 편차가 작을수록 집단간 편차가 클수록 큰 값이 됨**

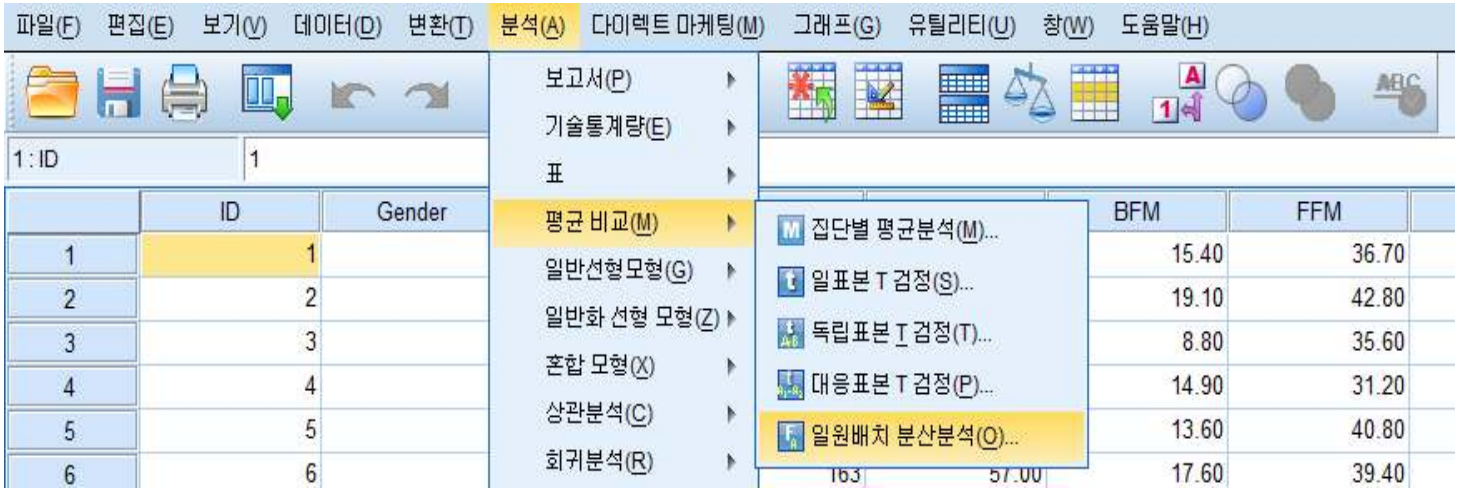


사후비교(post hoc comparisons)

- 귀무가설이 기각된 경우 어떤 집단간 차이가 있는지 2개씩 짝지어 차이를 분석함
- m개의 집단이면 ${}_mC_2 = m(m-1)/2$번 비교
- 집단 i와 집단 j를 비교하는 경우
  - $H_o : \mu_i = \mu_j$ $H_1 : \mu_i \neq \mu_j$
- 사후비교 방법
  - **Least Squate Difference method : 최소유의차, 검정력이 낮음**
  - **Bonferroni's method LSD : 방법을 보완**
  - **Scheffe's method : 사회과학에서 많이 사용**
  - **Tukey's method : 검정력이 높음**

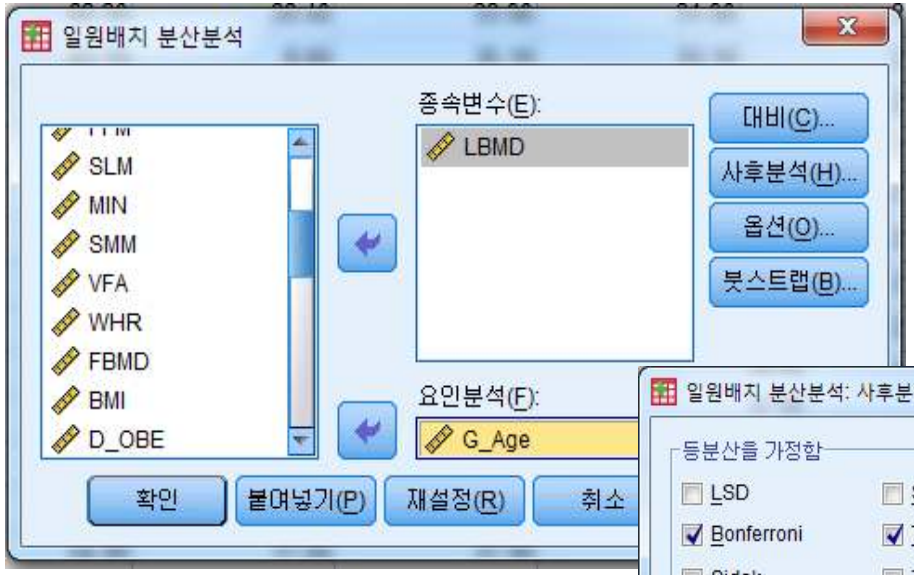
실습

- 연령집단 사이에 요추골밀도 T-값의 평균에 차이가 있는지 검정
- 범주형의 연령집단 변수는 "G_Age"로서
  - 연령집단을 다음 6가지로 분류하고 있다: 39세 이하, 40세~44세, 45세~49세, 50세~54세, 55세~59세, 그리고 60세 이상
  - 수치형의 요추골밀도 T-값을 나타내는 변수는 "LBMD"



실습

- 독립변수와 종속변수 지정



- 사후분석 지정

실습

- 옵션 지정

- 결과확인

기술통계

LBMD

	N	평균	표준편차	표준오차	평균에 대한 95% 신뢰구간		최소값	최대값
					하한값	상한값		
39	88	.400	1.0813	.1153	.171	.629	-2.3	3.0
40	76	.218	1.2748	.1462	-.073	.510	-2.6	5.2
45	160	.097	1.1152	.0882	-.077	.272	-2.4	4.0
50	231	-.433	1.2356	.0813	-.593	-.272	-2.9	4.7
55	115	-.857	1.2095	.1128	-1.081	-.634	-4.0	2.2
60	151	-1.042	1.3615	.1108	-1.261	-.823	-4.6	5.2
합계	821	-.351	1.3192	.0460	-.442	-.261	-4.6	5.2

실습

- 분산의 동질성 검정

분산의 동질성 검정

LBMD

Levene 통계량	df1	df2	유의확률
1.599	5	815	.158

- **ANOVA table**

분산분석

LBMD

	제곱합	df	평균 제곱	거짓	유의확률
집단-간	209.668	5	41.934	28.074	.000
집단-내	1217.353	815	1.494		
합계	1427.020	820			

- 다중비교

다중 비교

LBMD
Scheffe

(I) G_Age	(J) G_Age	평균차(I-J)	표준오차	유의확률	95% 신뢰구간	
					하한값	상한값
39	40	.1816	.1914	.970	-.457	.820
	45	.3025	.1622	.627	-.239	.844
	50	.8326*	.1531	.000	.322	1.343
	55	1.2574*	.1731	.000	.680	1.835
	60	1.4424*	.1639	.000	.896	1.989
40	39	-.1816	.1914	.970	-.820	.457
	45	.1209	.1703	.992	-.447	.689
	50	.6511*	.1616	.007	.112	1.190
	55	1.0758*	.1807	.000	.473	1.678
	60	1.2608*	.1719	.000	.687	1.834

보고서 작성

- **Mean Comparison of L-BMD T-score by Age Group**

L-BMD T-score	Age Group						*p*-value[1)]
	-39	40-44	45-49	50-54	55-59	60+	
Mean	0.400	0.218	0.097	-0.433	-0.857	-1.042	
SD	(1.0813)	(1.2748)	(1.1152)	(1.2356)	(1.2095)	(1.3615)	0.000
n	*n*=88	*n*=76	*n*=160	*n*=231	*n*=115	*n*=151	
T[2)]	a	a	a	b	b, c	c	

- **1) Statistical significances were tested by one-way analysis of variances among groups.**
- **2) The same letters indicate non-significant difference between groups (α=0.05) on Scheffe's multiple comparison test.**