경영데이터분석기초

- SPSS, Excel을 활용한 통계분석 -

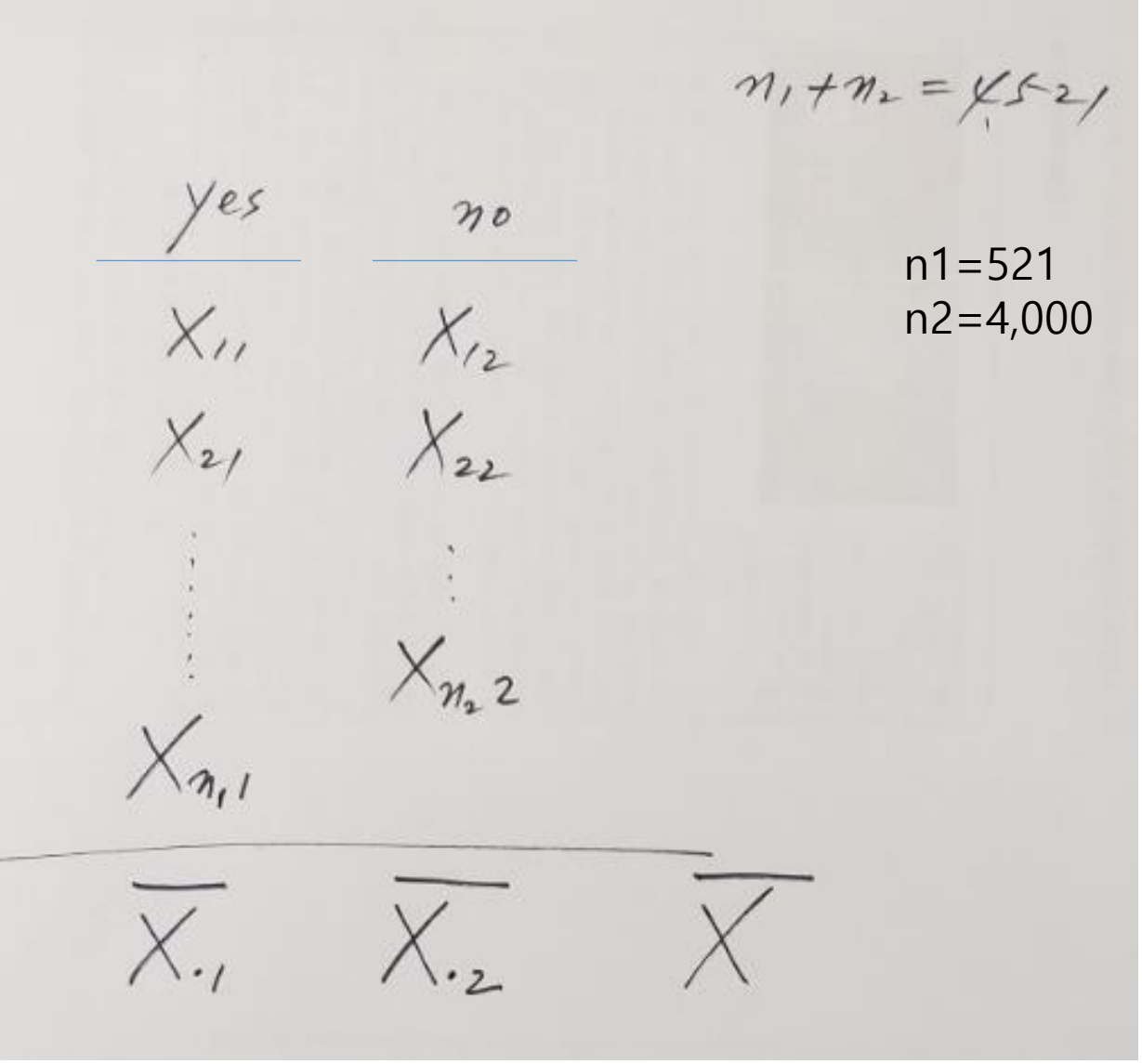
유 진 호

jhyoo@smu.ac.kr

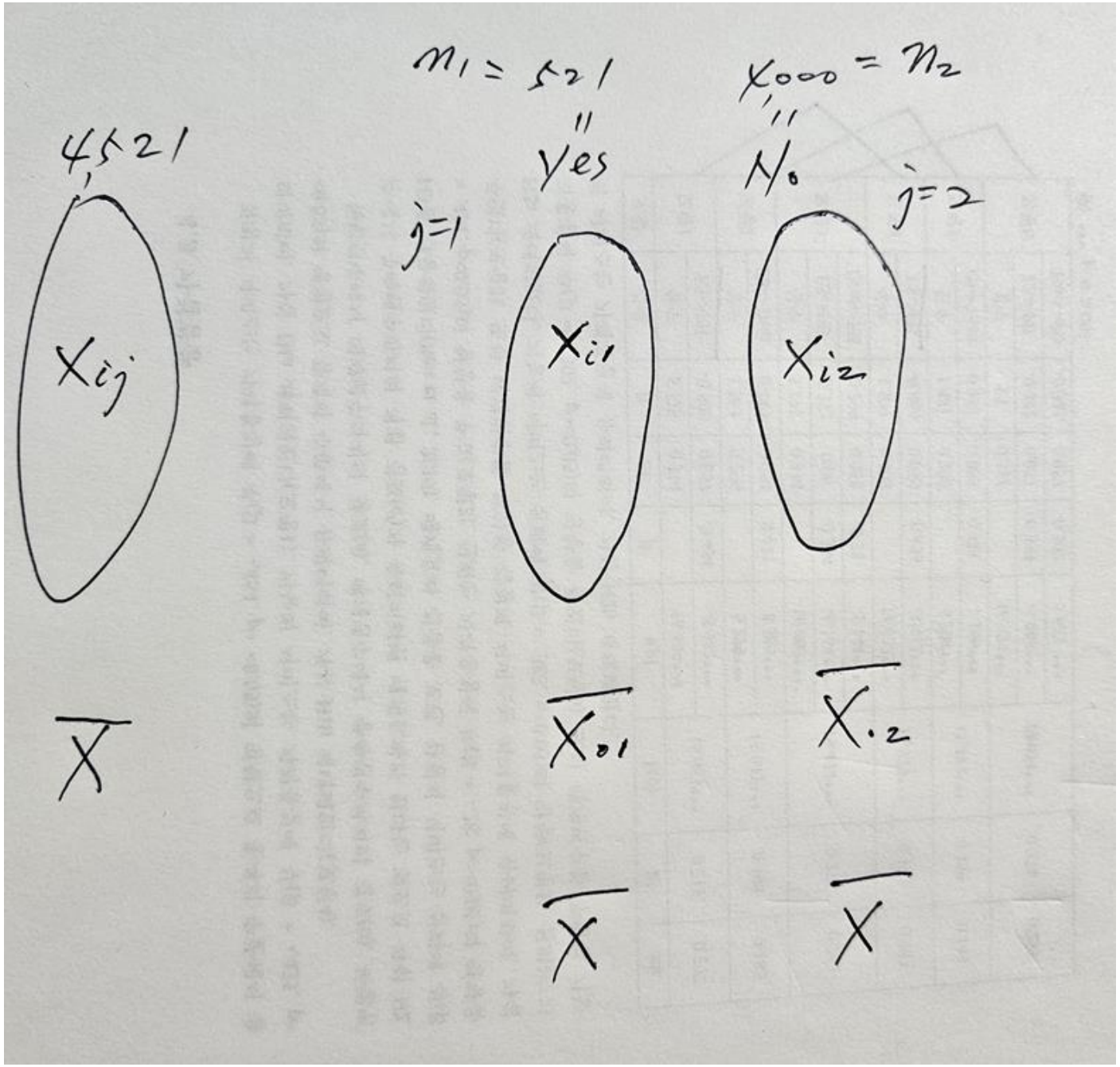
명목형 변수 vs. 연속형 변수 → 데이터탐색을 통한 기초통계분석, Box-plot 등 → 평균차이 검증

y duration

척도와 분석간의 관계		독립변수 X	
		범주형 자료	연속형 자료
종속변수 Y	범주형 자료	교차분석 (x^2검정)	로지스틱 회귀분석 판별분석
	연속형 자료	ANOVA(분산분석)	회귀분석



n1=521
n2=4,000

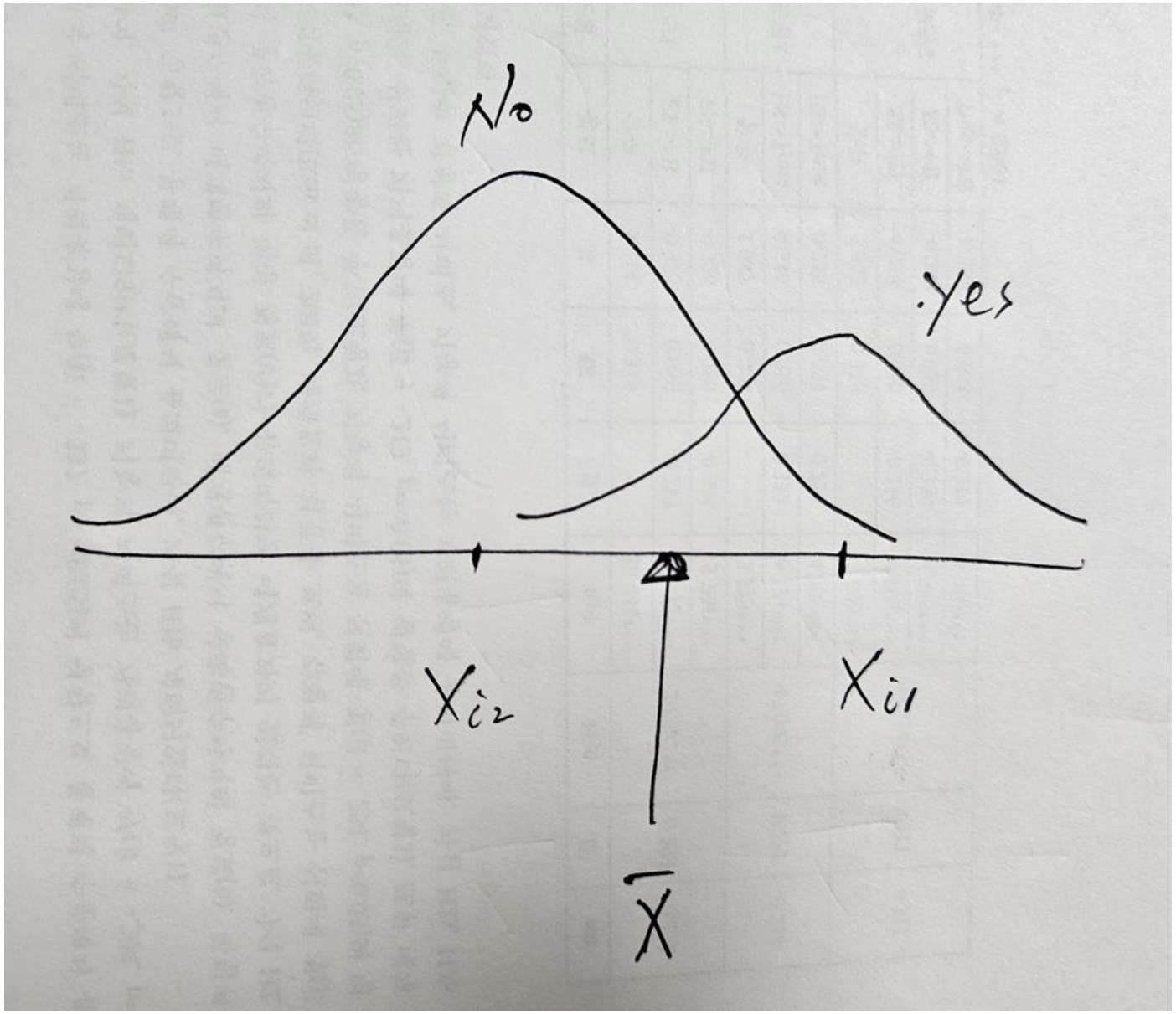


$\sum_j \sum_i (X_{ij} - \bar{X})^2$: 총변동 (총평균과의 차이)

$\sum_j \sum_i (X_{ij} - \bar{X}_{\cdot j})^2$: 집단내 변동 = 집단내 제곱합
(집단내 평균과의 차이)

$\sum_j n_j (\bar{X}_{\cdot j} - \bar{X})^2$: 집단간 변동 = 집단간 제곱합

집단 평균과 총평균과의 차이



$\sum_j \sum_i (X_{ij} - \bar{X})^2$: 총변동 (총평균과의 차이)

$\sum_j \sum_i (X_{ij} - \bar{X}_{\cdot j})^2$: 집단내 변동 = 집단내 제곱합
(집단내 평균과의 차이)

$\sum_j n_j (\bar{X}_{\cdot j} - \bar{X})^2$: 집단간 변동 = 집단간 제곱합

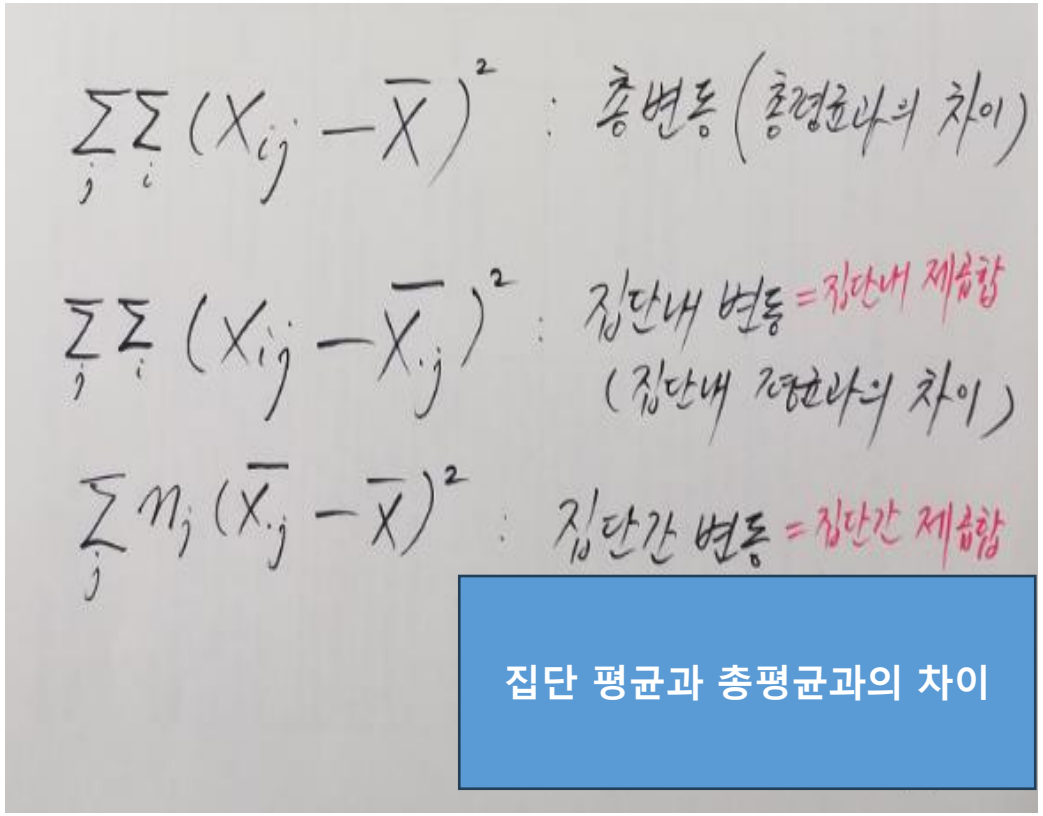
집단 평균과 총평균과의 차이

$$\text{검정통계량(F)} = \frac{\text{집단간 제곱합/집단간 자유도}}{\text{집단내 제곱합/집단내 자유도}}$$

평균차이 검정(검증)

- **총 변동 = 집단내 변동 + 집단간 변동**
- **집단내 변동**
  - 집단내 제곱합: 집단내 평균과의 차이를 의미
- **집단간 변동**
  - 집단간 제곱합: 집단 평균과 총 평균과의 차이를 의미

$$\text{검정통계량(F)} = \frac{\text{집단간 제곱합/집단간 자유도}}{\text{집단내 제곱합/집단내 자유도}}$$



j= 1(yes), 2(no)

$n_1 + n_2 = 4521$

Y = yes / no

X_{11} X_{21} ⋮ $X_{n_1 1}$

X_{12} X_{22} ⋮ $X_{n_2 2}$

집단내 평균

$\bar{X}_{\cdot 1}$ $\bar{X}_{\cdot 2}$ $\bar{X}$

전체평균 $\left(X_{11} + X_{12} + \cdots + \right)$ / 4521

$\dfrac{X_{11} + \cdots + X_{n_1 1}}{n_1}$

$\dfrac{X_{12} + \cdots + X_{n_2 2}}{n_2}$

4521 − 1 = 4520

$\sum_j \sum_i (X_{ij} - \bar{X})^2$: 총변동 (총평균과의 차이)^2

$\sum_j \sum_i (X_{ij} - \bar{X}_{\cdot j})^2$: 집단내 변동 = 집단내 제곱합 (집단내 평균과의 차이)^2

$\sum_{j=1}^{2} n_j (\bar{X}_{\cdot j} - \bar{X})^2$: 집단간 변동 = 집단간 제곱합

2 1

집단 평균과 총평균과의 차이

$n_1 - 1$, $n_2 - 1$ $= (n_1 + n_2 - 2)$

$= 4521 - 2 = 4519$

2 − 1 = ①

평균차이 검정은 두 그룹(집단)의 평균을 비교해서 이들의 차이가 모집단에도 있다고 할 수 있는지를 검정하는 것이다. 검정대상이 되는 개체를 두 그룹(집단)으로 구분하여 측정한 다음, 이들의 평균을 비교해서 검정하게 된다.

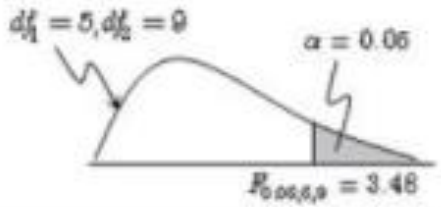
귀무가설 : 두 그룹(집단)의 모평균에 차이가 없다.(즉, 동일하다)

대립가설 : 두 그룹(집단)의 모평균에 차이가 있다.(즉, 동일하지 않다)

평균차이의 검정을 3 집단 이상으로 확장하면 F 분포를 이용하여 검정한다.

$$\text{검정통계량}(F) = \frac{\text{집단간 제곱합/집단간 자유도}}{\text{집단내 제곱합/집단내 자유도}}$$

F분포표



(분모의 자유도) df_2	(분자의 자유도) df_1									
	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54

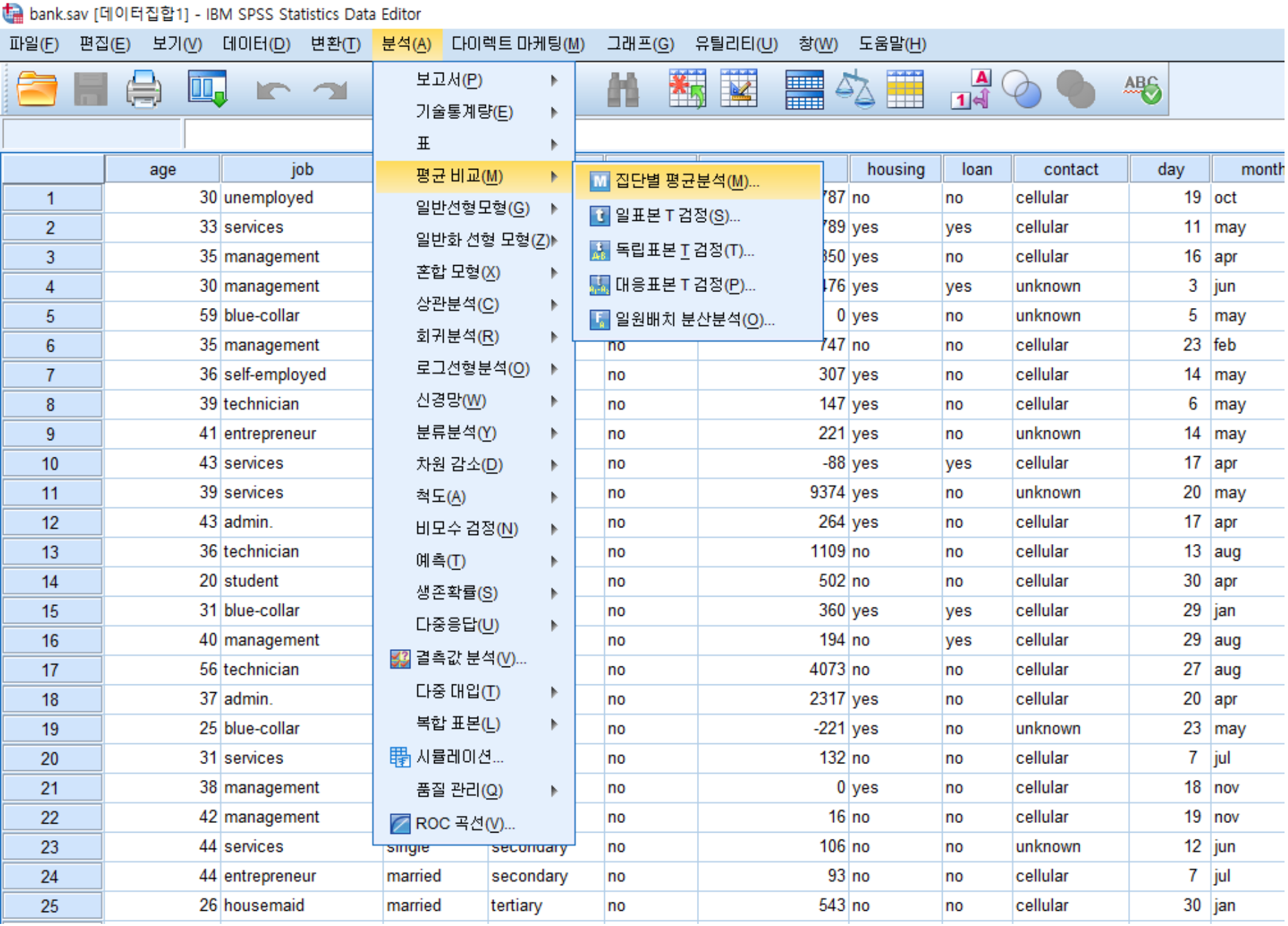
=fdist(x,df1,df2)

(예)

=fdist(3.48,5,9)
➔ 0.05

확률값=F.Dist.RT(x , 1, 4519)

임계값=F.INV.RT(0.05, 1, 4519)



연속형 변수와 타겟변수와 연관성 분석
(타겟변수에 영향을 줄 수 있는 연속형 변수 찾기)
.Means (평균 차이 검증)

결과해석하기

- 상관분석
  - Age와 balance의 상관계수는 ###, 유의확률은 ###이므로 유의수준 0.05에서 상관성이 존재하다. 다만, 상관계수값이 작으므로 아주 작은 상관관계가 존재한다고 할 수 있다.

- 교차분석(CrossTab)
  - 카이제곱(X^2) 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 [Y 집단]별로 [결혼상태]는 서로 다르다(서로 연관성이 있다)고 할 수 있다.

- 평균차이분석(Means)
  - F통계량 값이 866.5, 유의확률은 0.000 이므로 유의수준 0.05에서 [Y 집단]별로 [duration]은 통계적으로 유의한 차이가 있다고 할 수 있다.

- 회귀분석
  - 분산분석표에 따라 F 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 회귀모형이 통계적으로 유의하다고 할 수 있다.
  - t 통계량값이 ###, 유의확률은 ###이므로 유의수준 0.05에서 회귀계수는 0이 아니다. 즉, [나이]는 [balance]에 영향을 준다(나이에 따라 balace는 달라진다)고 할 수 있다.

과제

1. SPSS로 Y집단별 duration 평균차이에 대한 검증을 하고, 그 결과를 엑셀에서 계산한 값과 비교하기(excel 자료 제출하기)

2. SPSS로 Y집단과 6개 연속형 변수들과의 평균차이에 대한 검증을 각각 실시하고 그 결과를 excel파일로 export(내보내기)하여 통계 결과에 대한 해석을 달기

   * 6개의 연속형/수치형 변수: age balance duration campaign pdays previous,