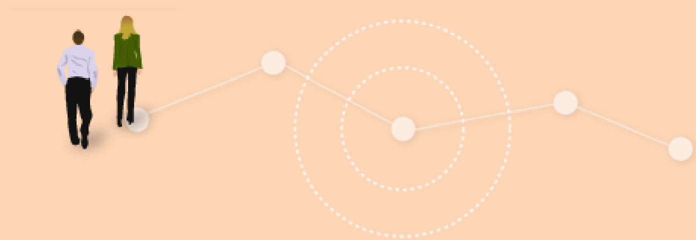
2 장

목 차

01 데이터의 수집

:: Keywords 변수 | 데이터 | 척도



변수와 데이터

■ 변수

어떠한 대응관계로 변화하는 수(數), 혹은 함수관계로 대응하며

주어진 범위 안에서 변화하는 수(數)

→ 변수는 데이터로 구성되고, 데이터를 근거로 변수의 특성을 파악

■ 데이터

조사의 목적에 맞는 변수를 기반으로, 표본으로부터 수집된 자료

→ 보통 사회과학 분야에서 통계조사를 할 때는 표본의 특징이나

특성을 표현하기 위해 단일자료를 수집,

핵심적 연구나 조사를 목적으로 할 때는 다중자료를 수집

척도

데이터는 그 성격에 따라 크게 범주형 척도와 연속형 척도의 두 가지로 구분

인문/사회과학에서는 대부분 설문지를 통해 데이터를 수집할 때 활용

■ 범주형 척도

범주형 척도(categorical scale)는 데이터들을 구분지어 나눌 수 있는 척도

→ 명목척도와 서열척도로 나뉜다.

■ 연속형 척도

연속형 척도(continuous scale)는 연결된 속성의 데이터를 조사의 목적에 맞게 구분한 척도

→ 등간척도와 비율척도로 나뉜다.

범주형 척도

■ 명목척도

명목척도(nominal scale)는 수(數) 또는 순서의 개념과는 상관없이 이름만 붙여지는 척도

예 설문 문항의 보기는 대부분 '① 남자 ② 여자'와 같이 주어지며, 응답자는 이에 대해 '①' 혹은 '②'로 답한다. 이때 '남자'나 '여자'는 '1'과 '2'라는 숫자 또는 연산과는 아무런 연관이 없다. 즉 '남자 + 남자'를 수식으로 나타내자면 '1 + 1 = 2'이지만, 남자 두 명을 더한다하여 '여자'가 될 수는 없다는 의미이다.

범주형 척도

■ 서열척도 (순서척도)

명목척도와 유사하게 숫자 혹은 연산과는 연관이 없지만,

순서(서열)을 구분할 수 있는 척도

> 예 마라톤 경기 결과처럼 '1등', '2등', '3등' 혹은 '금메달', '은메달', '동메달'을 구분할 때, '1등 + 2등'이 '3등' 혹은 '은메달 + 동메달'이 '금메달'이 될 수는 없다. 이때 주의해야 할 것은 '2−1 = 1'이고 '3−2 = 1'로서 모두 1이라는 차이가 나지만, 1등과 2등 사이의 차이가 2등과 3등 사이의 차이(혹은 거리)와 같다고 할 수 없다는 것이다.

연속형 척도

■ 등간척도

등간척도(interval scale)는 명목척도나 서열척도와 달리, 측정된 자료들 간에 더하기와 빼기가 가능한 척도

예 ‘섭씨 영상 15도’를 생각해보자. 이는 ‘0도보다 15도만큼 높은 온도’이다. 여기서 0도라는 것은 영상과 영하의 구분점이 되는 온도이지 온도 자체가 없는 무(無)를 의미하는 ‘절대 0’의 개념이 아니다.

■ 비율척도

비율척도(ratio scale)는 등간척도의 성질과 함께 무(無)의 개념인 0값도 가지는 척도

예 ‘길이’, ‘무게’, ‘부피’, ‘경력’ 등과 같이 다양한 기준을 비율척도로 활용할 수 있다.

연속형 척도

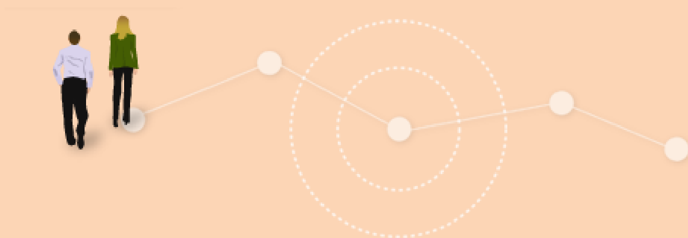
참고 등간척도와 비율척도의 계산 개념

등간척도와 비율척도를 구분하는 기준은 계산 가능 여부, 또 절대 0의 개념이 존재하는지의 여부이다. 등간척도의 '매우 만족 = 5', '만족 = 4', '보통 = 3', '불만족 = 2', '매우 불만족 = 1'로 설문을 했을 때, '매우 만족 = 5' – '만족 = 4' = '매우 불만족 = 1'이라는 계산이 가능한지 의문이 생길 수 있다. 계산이 가능하다고 했기 때문에 틀린 것 같지 않지만, 이 부분은 단순히 산술적인 접근으로 나타내면 오류를 범하게 된다.

애초에 척도를 구분하는 기준은 숫자 개념이 아니다. 크게 범주형 척도와 연속형 척도로 구분되어 있는 상황에서 더욱 세분화하기 위해 '숫자의 개념을 어느 정도 가지고 있는가'라는 기준을 적용한 것이다. 그러므로 '5 – 4 = 1'의 산술적인 접근이 아니라 '매우 만족' – '만족' = '나머지 만족'으로 생각해야 한다.

02 데이터의 표현 방법

:: Keywords 도수분포표 | 그래프



도수분포표

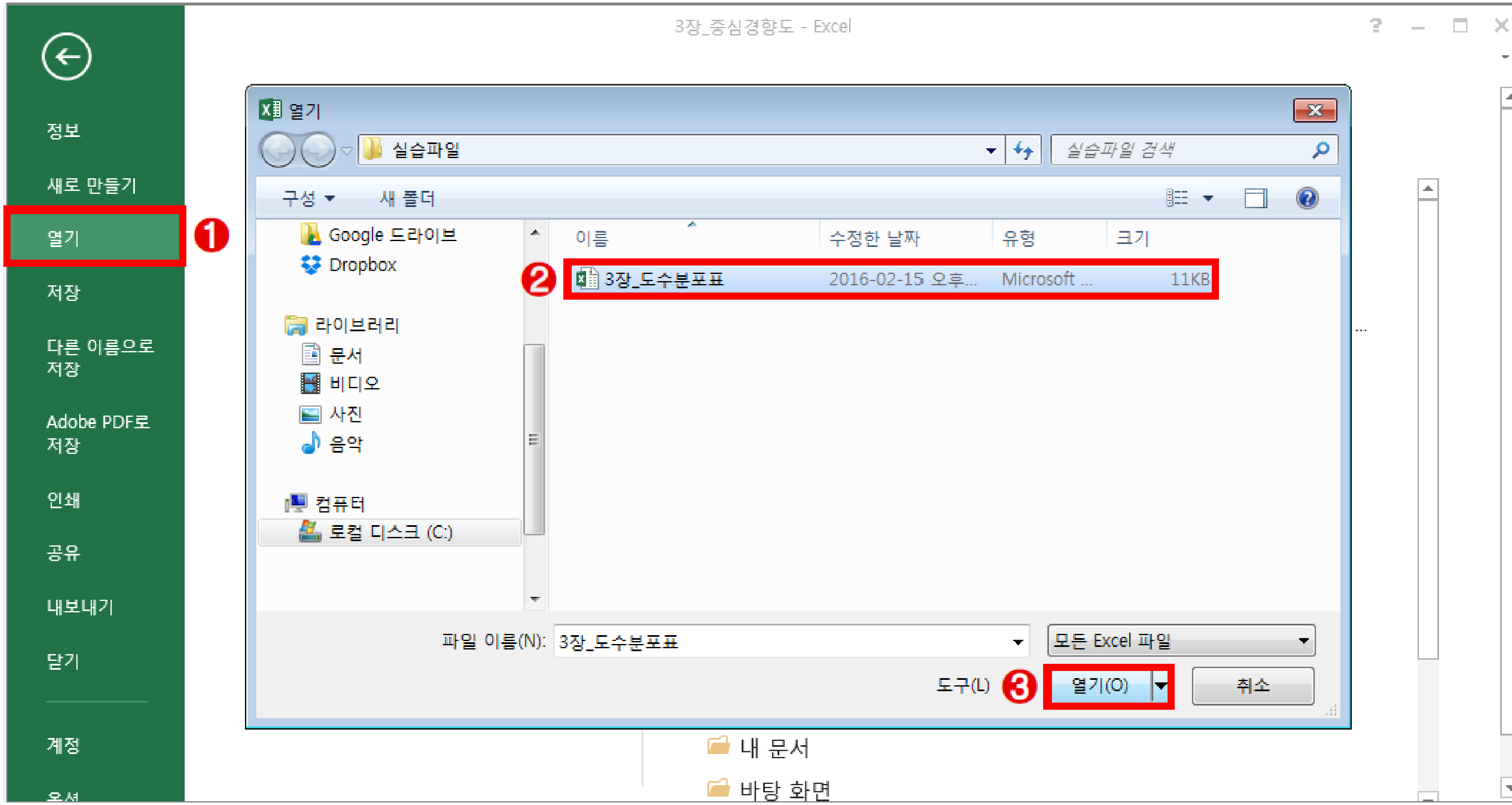
■ 도수분포표

도수분포표(frequency distribution table)는 수집된 각각의 데이터에 대한 개수를 정리한 표

예제 3-1 **도수분포표 출력** 준비파일 | 3장_도수분포표.xlsx

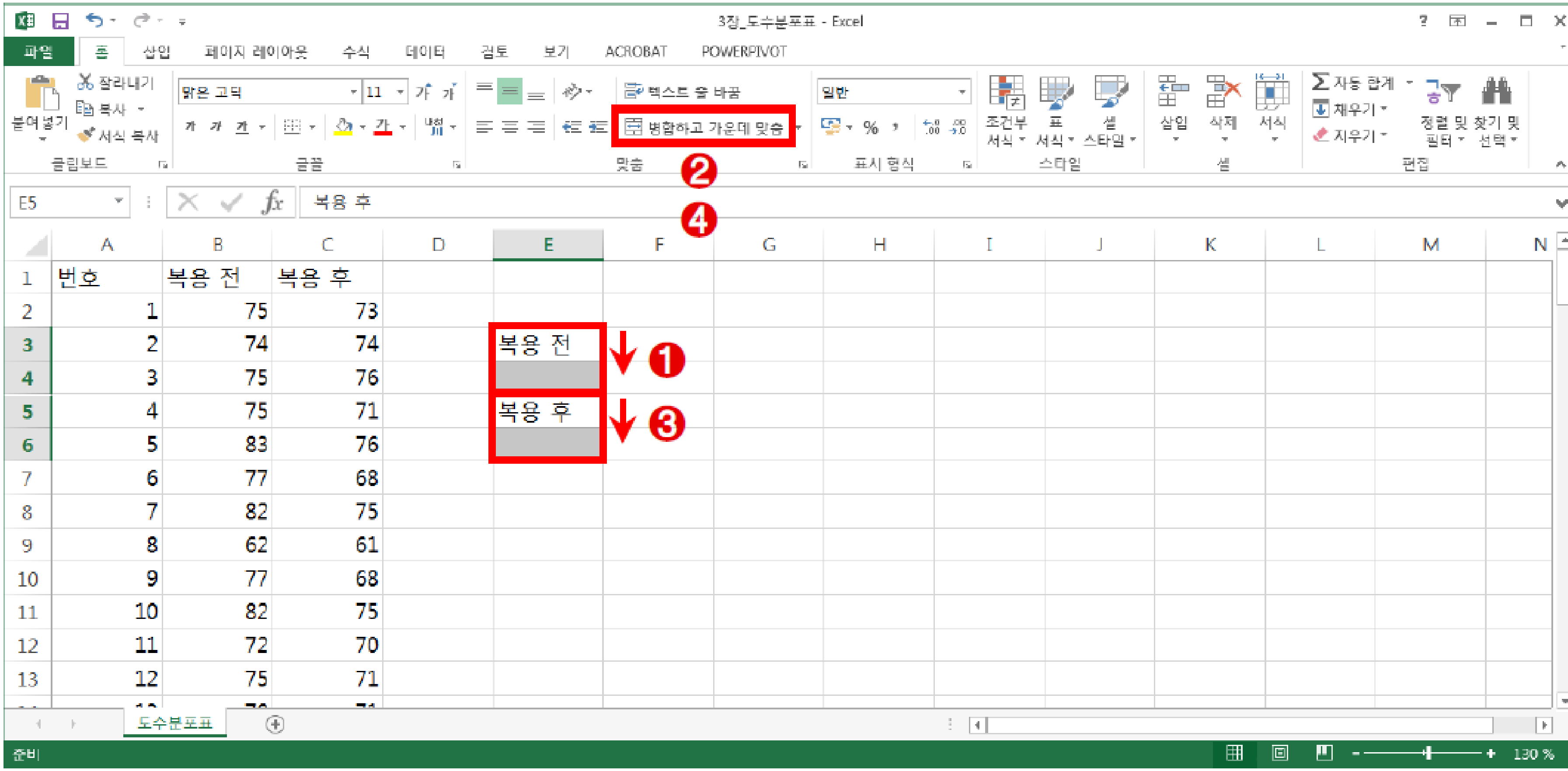
'3장_도수분포표.xlsx' 파일은 A제약회사에서 100명을 대상으로 다이어트 신약의 복용 전과 후의 체중 변화를 조사한 자료이다. 주어진 데이터의 최대값과 최소값을 이용하여 계급 구간을 설정하고 도수분포표를 출력하라.

도수분포표

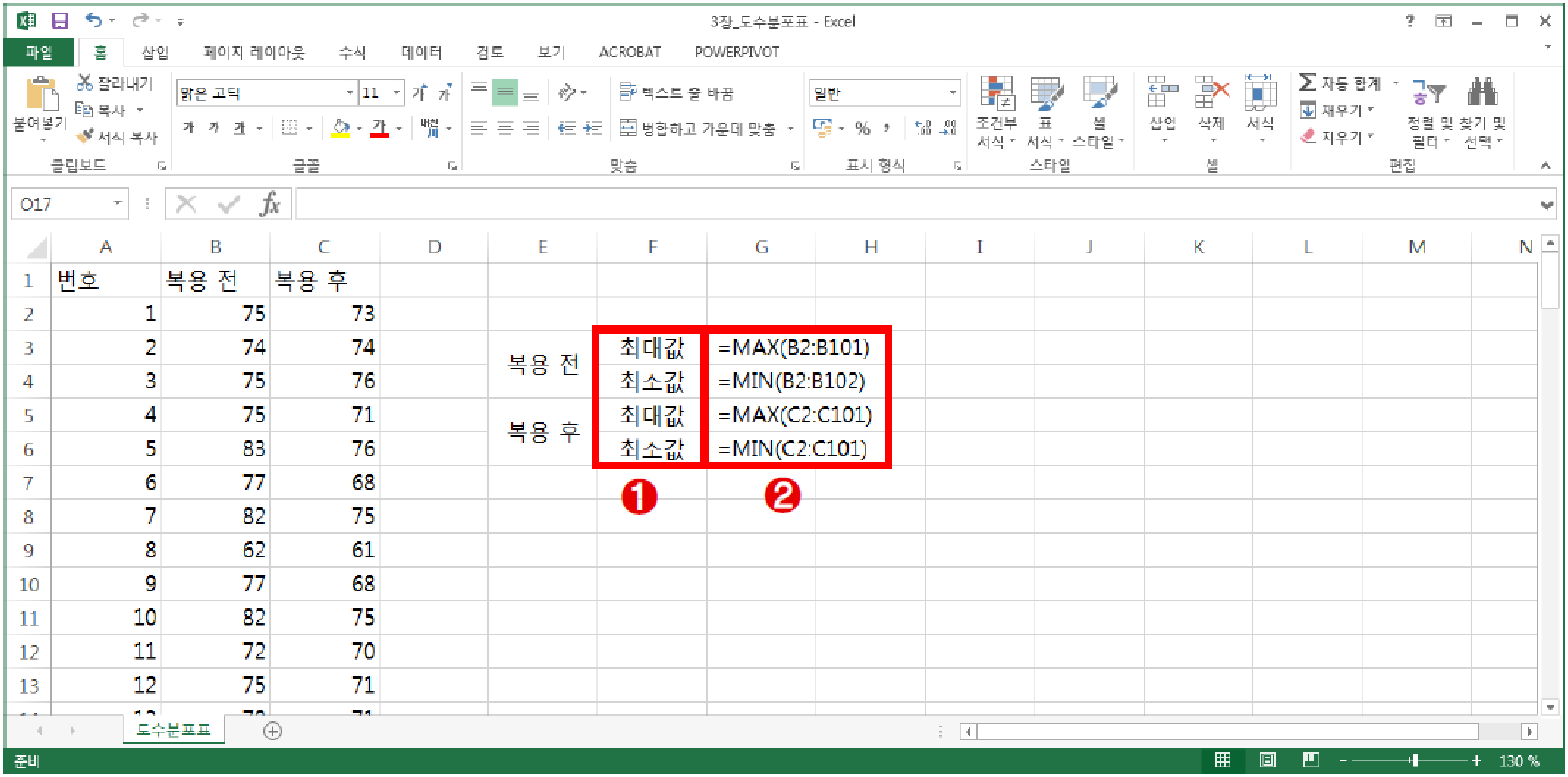


3장_중심경향도 - Excel
정보
새로 만들기
열기
저장
다른 이름으로 저장
Adobe PDF로 저장
인쇄
공유
내보내기
닫기
계정
열기
실습파일
실습파일 검색
구성
새 폴더
Google 드라이브
Dropbox
라이브러리
문서
비디오
사진
음악
컴퓨터
로컬 디스크 (C:)
이름
수정한 날짜
유형
크기
3장_도수분포표
2016-02-15 오후...
Microsoft ...
11KB
파일 이름(N):
3장_도수분포표
모든 Excel 파일
도구(L)
열기(O)
취소
내 문서
바탕 화면

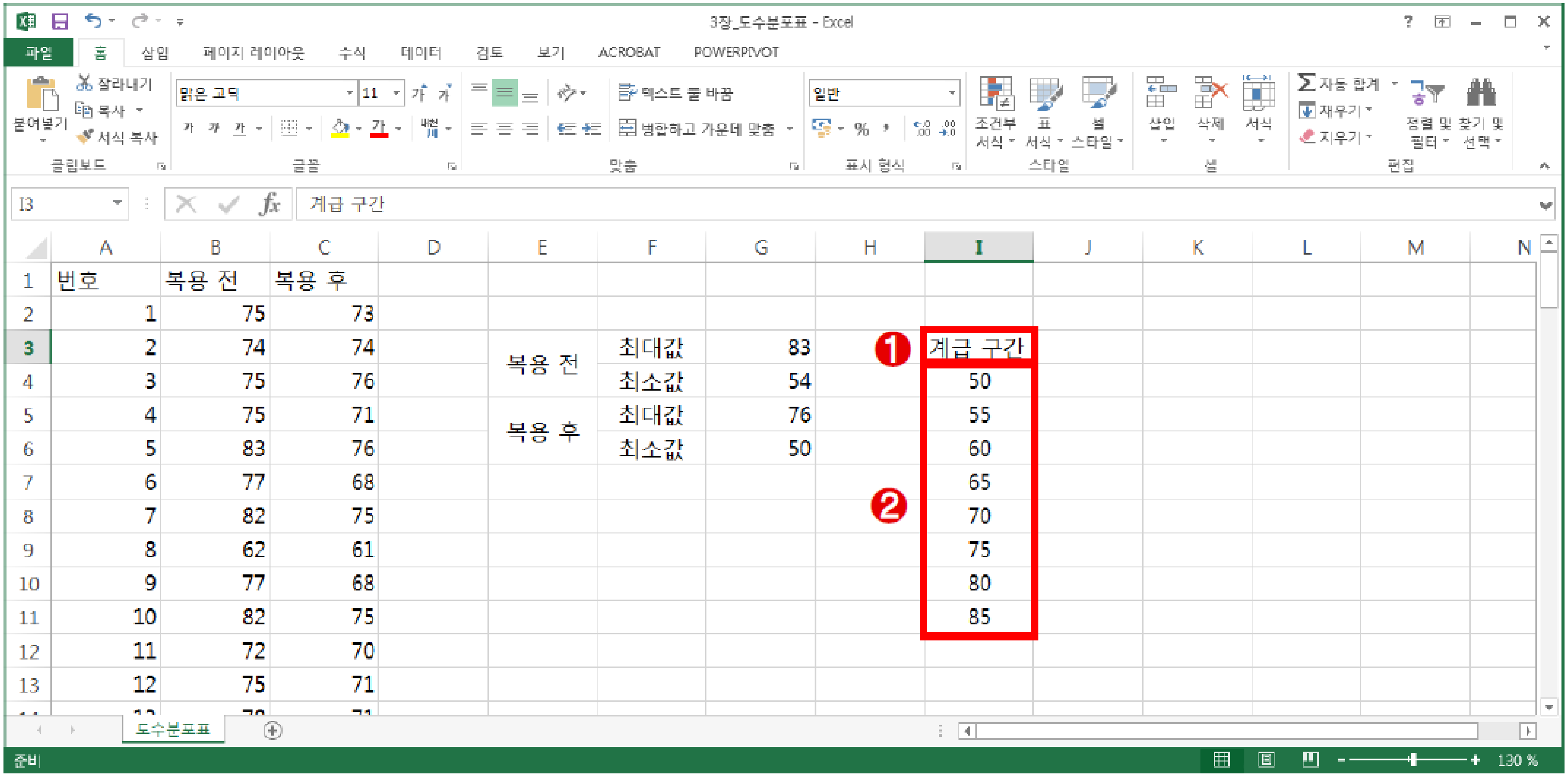
도수분포표



도수분포표

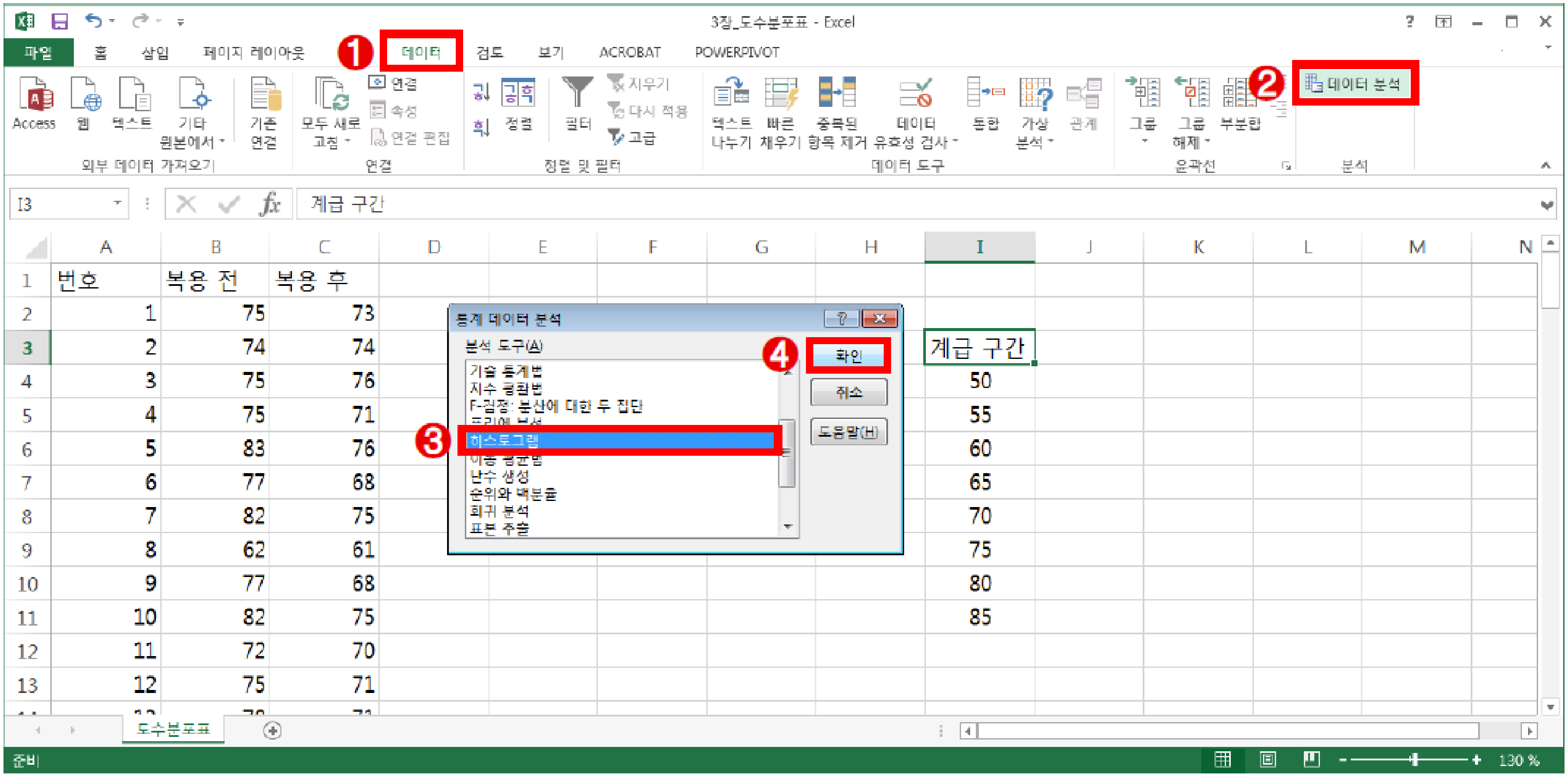

3장_도수분포표 - Excel
번호 복용 전 복용 후
1 75 73
2 74 74
3 75 76
4 75 71
5 83 76
6 77 68
7 82 75
8 62 61
9 77 68
10 82 75
11 72 70
12 75 71
복용 전
최대값 =MAX(B2:B101)
최소값 =MIN(B2:B102)
복용 후
최대값 =MAX(C2:C101)
최소값 =MIN(C2:C101)
❶
❷
도수분포표

도수분포표

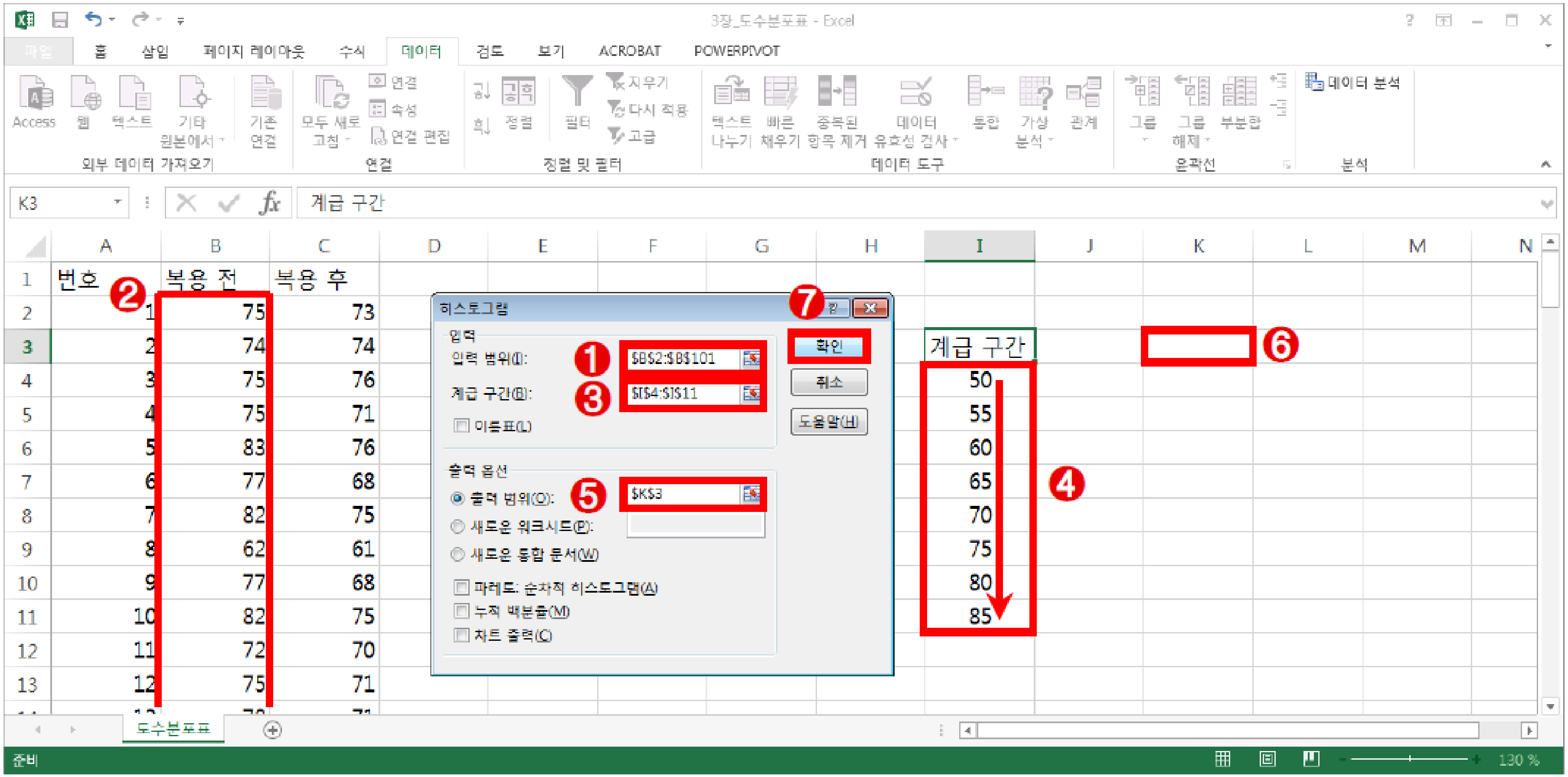


3장_도수분포표 - Excel
파일 홈 삽입 페이지 레이아웃 수식 데이터 검토 보기 ACROBAT POWERPIVOT
I3 계급 구간
번호 복용 전 복용 후
복용 전 최대값 83
최소값 54
복용 후 최대값 76
최소값 50
❶ 계급 구간
❷ 50 55 60 65 70 75 80 85
1 75 73
2 74 74
3 75 76
4 75 71
5 83 76
6 77 68
7 82 75
8 62 61
9 77 68
10 82 75
11 72 70
12 75 71
도수분포표
준비
130 %

도수분포표

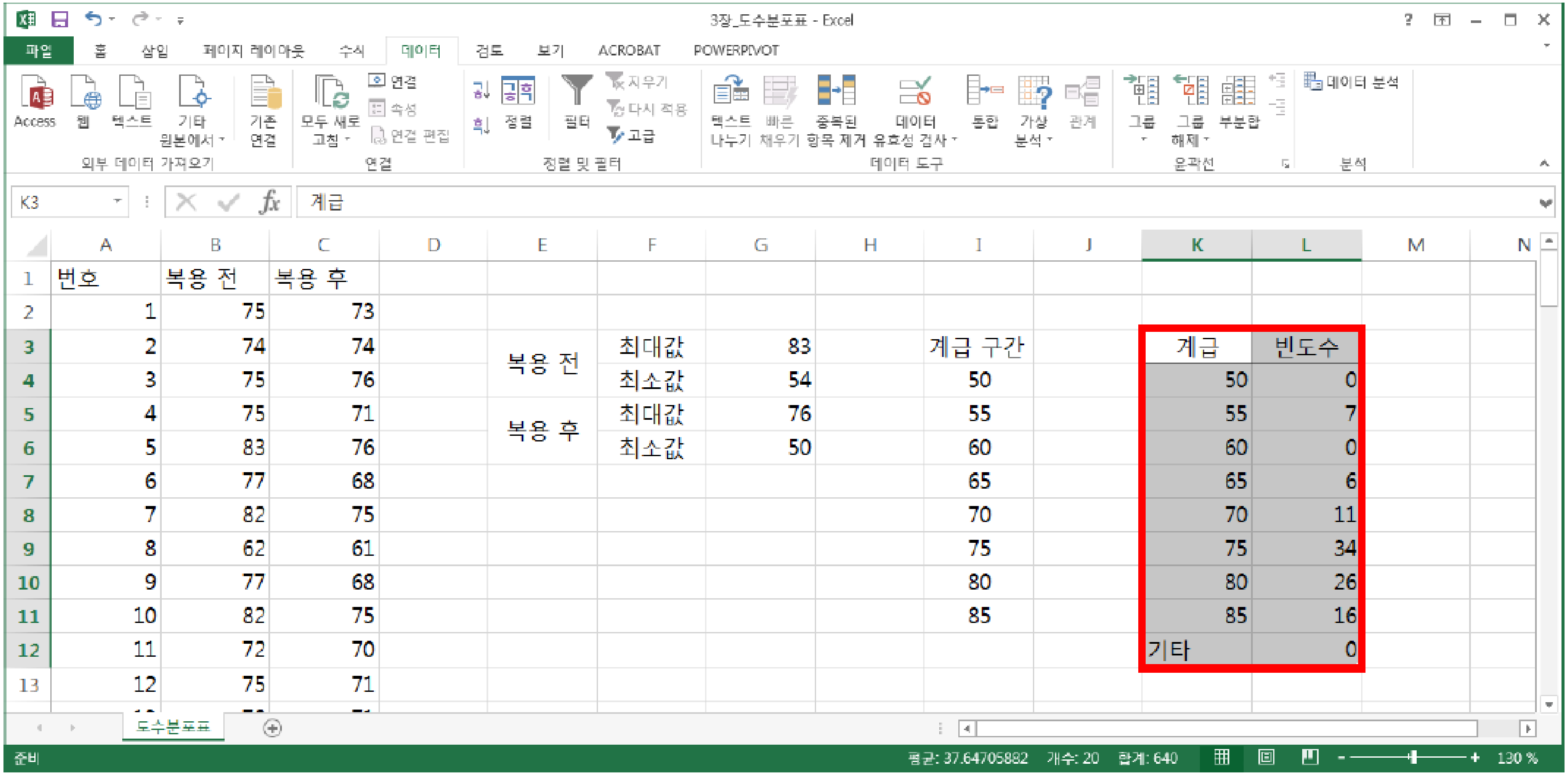

3장_도수분포표 - Excel
파일 홈 삽입 페이지 레이아웃 데이터 검토 보기 ACROBAT POWERPIVOT
데이터 분석
통계 데이터 분석
분석 도구(A)
기술 통계법
지수 평활법
F-검정: 분산에 대한 두 집단
히스토그램
난수 생성
순위와 백분율
회귀 분석
표본 추출
확인
취소
도움말(H)
번호
복용 전
복용 후
계급 구간
50
55
60
65
70
75
80
85
도수분포표
준비

도수분포표



히스토그램
입력
입력 범위(I):
$B$2:$B$101
계급 구간(B):
$I$4:$I$11
이름표(L)
출력 옵션
출력 범위(O):
$K$3
새로운 워크시트(P):
새로운 통합 문서(W)
파레토: 순차적 히스토그램(A)
누적 백분율(M)
차트 출력(C)
확인
취소
도움말(H)
계급 구간
번호
복용 전
복용 후

도수분포표(완성)



	A	B	C	D	E	F	G	H	I	J	K	L
1	번호	복용 전	복용 후									
2	1	75	73									
3	2	74	74		복용 전	최대값	83		계급 구간		계급	빈도수
4	3	75	76			최소값	54		50		50	0
5	4	75	71		복용 후	최대값	76		55		55	7
6	5	83	76			최소값	50		60		60	0
7	6	77	68						65		65	6
8	7	82	75						70		70	11
9	8	62	61						75		75	34
10	9	77	68						80		80	26
11	10	82	75						85		85	16
12	11	72	70								기타	0
13	12	75	71									

그래프

■ 그래프

도수분포표의 단점

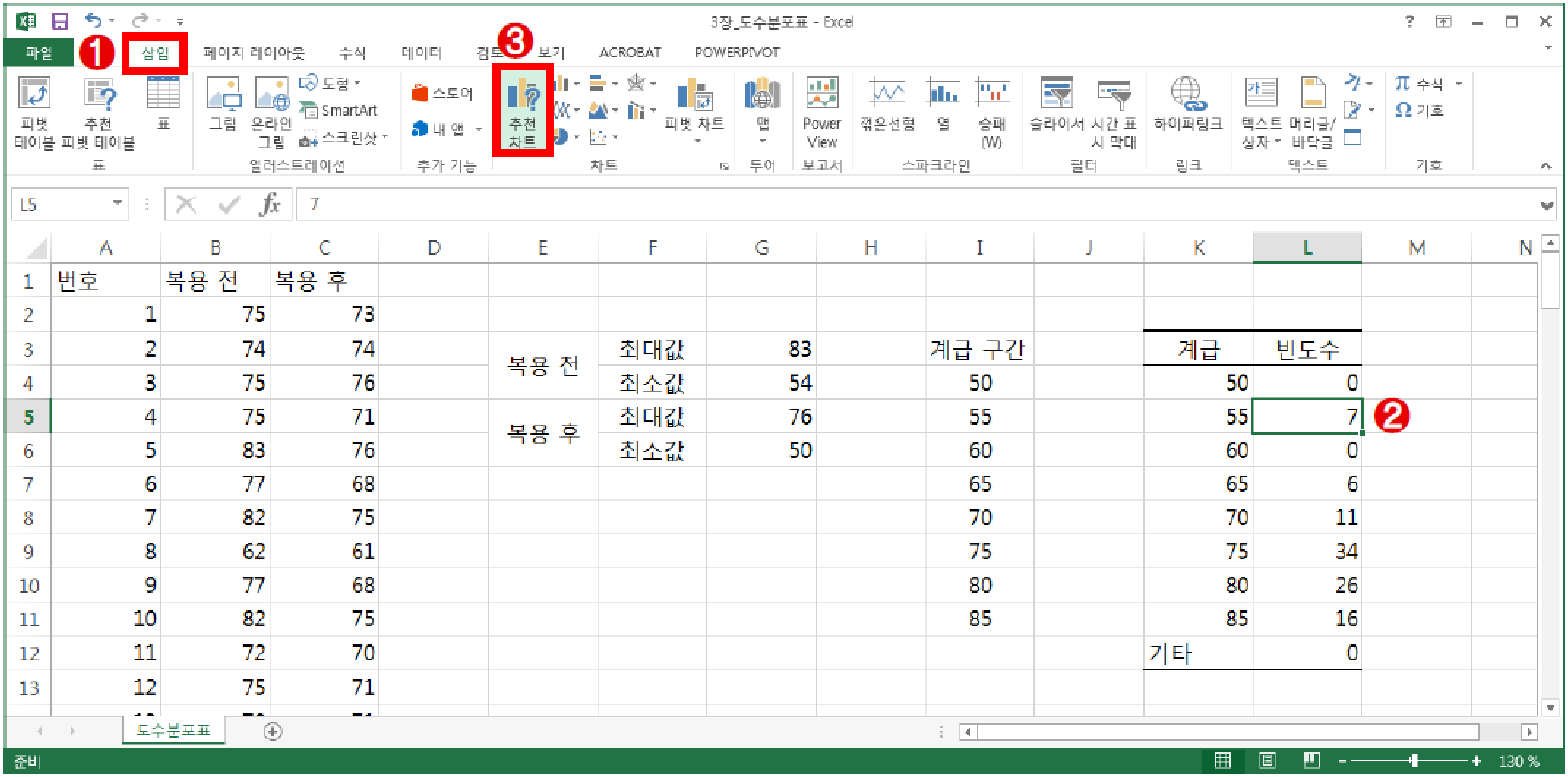
: 직관적으로 인식되는 형태가 아니어서 일일이 숫자를 비교해야 함

→ 그래프는 숫자를 일일이 살펴보지 않아도 크기나 형태 등을 바로 비교 가능

예제 3-2 그래프 출력 준비파일 | 3장_도수분포표.xlsx

[예제 3-1]에서 구한 도수분포표를 이용하여 막대그래프를 출력하라.

그래프


3장_도수분포표 - Excel
파일
삽입
페이지 레이아웃
수식
데이터
검토
보기
ACROBAT
POWERPIVOT
추천 차트
L5
7
번호
복용 전
복용 후
최대값
최소값
계급 구간
계급
빈도수
기타
도수분포표
준비
130 %

그래프

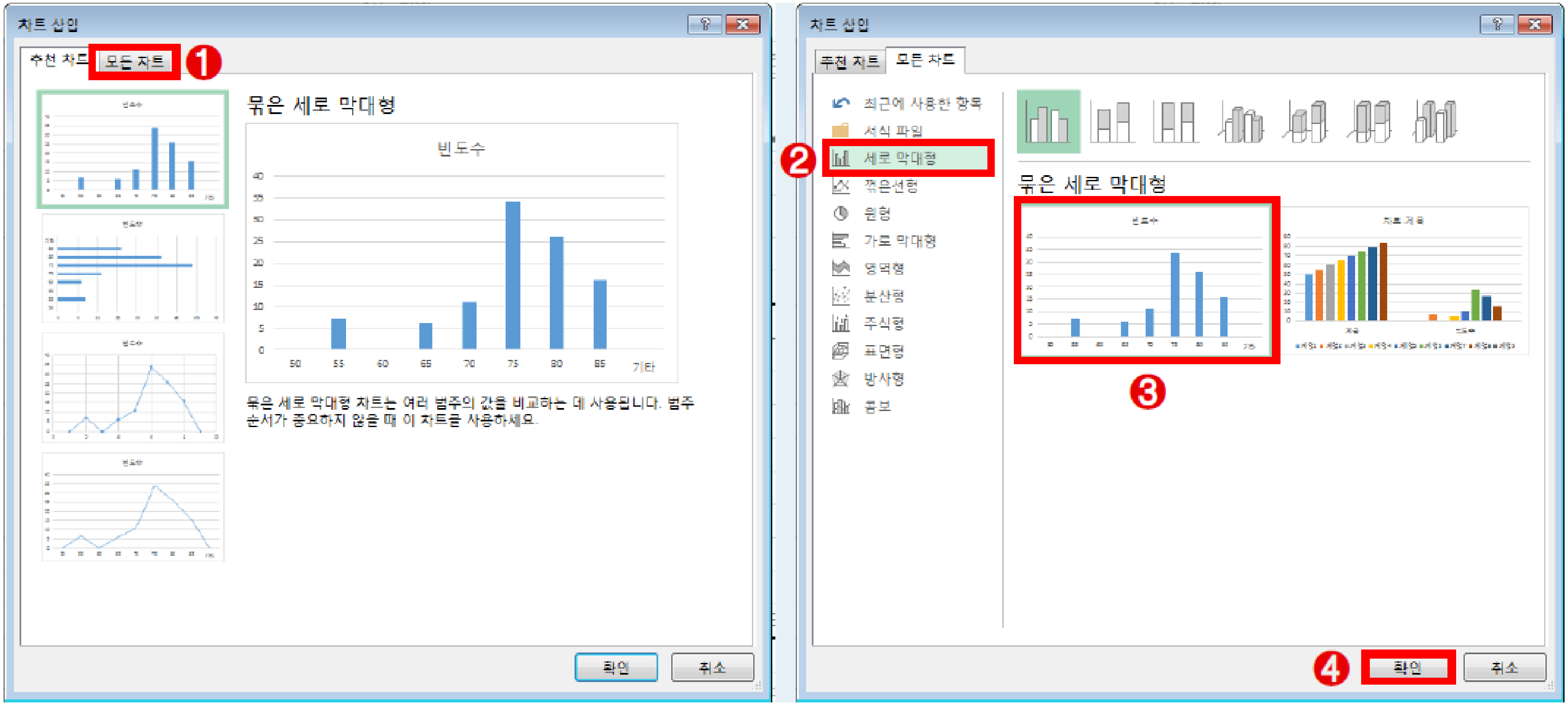
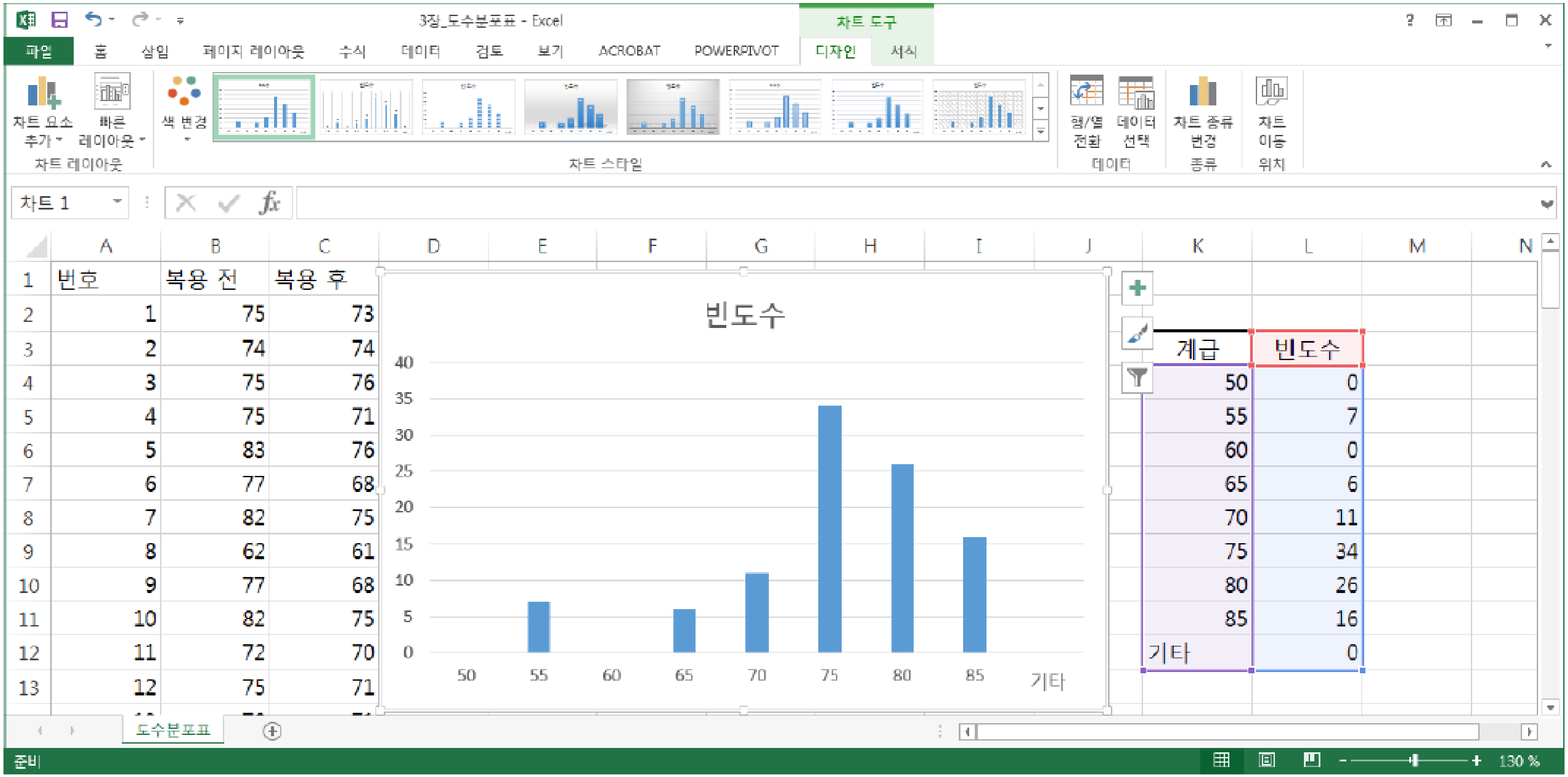


차트 삽입
추천 차트
모든 차트
❶
묶은 세로 막대형
빈도수
묶은 세로 막대형 차트는 여러 범주의 값을 비교하는 데 사용됩니다. 범주
순서가 중요하지 않을 때 이 차트를 사용하세요.
확인
취소
차트 삽입
추천 차트
모든 차트
최근에 사용한 항목
서식 파일
❷
세로 막대형
꺾은선형
원형
가로 막대형
영역형
분산형
주식형
표면형
방사형
콤보
묶은 세로 막대형
❸
❹
확인
취소

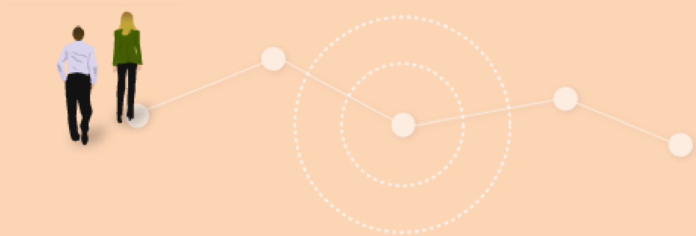
그래프 (완성)



3장_도수분포표 - Excel
차트 도구
파일 홈 삽입 페이지 레이아웃 수식 데이터 검토 보기 ACROBAT POWERPIVOT 디자인 서식
차트 요소 추가
빠른 레이아웃
색 변경
차트 레이아웃
차트 스타일
행/열 전환
데이터 선택
데이터
차트 종류 변경
종류
차트 이동
위치
차트 1
번호 복용 전 복용 후
빈도수
계급 빈도수
50 0
55 7
60 0
65 6
70 11
75 34
80 26
85 16
기타 0
도수분포표
준비
130 %

03 기초 통계량

:: Keywords 평균 | 중간값 | 최빈값 | 분산 | 표준편차



중심경향도

중심경향도(measure of central tendency)란 데이터들을 종합하여 그 중심을 이루는 값이 어느 정도가 될지를 구한 것

Ex. Q. 학교까지 가는 데 어느 정도의 시간이 걸리는가?

A. 30분 정도 소요됩니다.

A. 30분~40분 정도 소요됩니다.

→ 통학 시간 전체라는 집단 특성을 대표적으로 표현하는 값에는 평균, 중간값, 최빈값 등이 있음

중심경향도

■ 평균

평균(mean)은 통계에서 가장 많이 활용되는 중심경향도

→ 모든 통계분석에서 사용되며 표본의 특성을 제시할 때 가장 먼저 사용되는 수치

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \cdots + x_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

참고 산술평균, 기하평균, 조화평균

a와 b가 실수일 때

$$(\text{산술평균}) = \frac{a+b}{2},\ (\text{기하평균}) = \sqrt{ab},\ (\text{조화평균}) = \frac{2ab}{a+b}$$

이며, 이들 사이에는 다음의 관계가 성립한다.

$$\frac{a+b}{2} \geq \sqrt{ab} \geq \frac{2ab}{a+b}$$

$$(\text{산술평균}) \geq (\text{기하평균}) \geq (\text{조화평균})$$

중심경향도

■ 중간값

중간값(median)은 관측된 자료의 편중과는 상관없이 가장 작은 값에서 가장 큰 값까지 정렬했을 때 그 가운데 위치한 값

■ 최빈값

최빈값(mode)은 표본에서 가장 많이 나타나는 관측치

→ 여러 번 확인된다는 특성으로 중심경향도에 있으나, 최소 부분과 최대 부분으로 쏠림 현상이 나타날 수도 있기 때문에 특별히 필요성이 없는 한 잘 사용되지 않음

중심경향도

예제 3-3 평균, 중간값, 최빈값 계산 준비파일 | 3장_중심경향도.xlsx

평소 공부에 관심이 없는 A학생의 부모님은 자녀의 성적 향상에 대한 의욕을 높여주기 위해, 이번 중간고사 성적이 반에서 중간 이상이면 선물을 사주겠다고 약속했다. A학생의 반은 6명이고, 성적이 각각 10, 40, 70, 85, 85, 100점이었다. A학생이 70점을 받았을 때, 과연 중간 이상의 성적을 받은 것인지 평균, 중간값, 최빈값을 기준으로 설명하라.

중심경향도 (풀이)

■ 평균

$$\frac{10 + 40 + 70 + 85 + 85 + 100}{6} = 65$$

→ A학생의 점수는 70점, 반 평균 점수는 65점이다.
따라서 평균 기준으로는 A학생의 점수가 평균보다 5점이나 더 높으므로
부모님으로부터 선물을 받을 수 있다.

중심경향도 (풀이)

■ 중간값

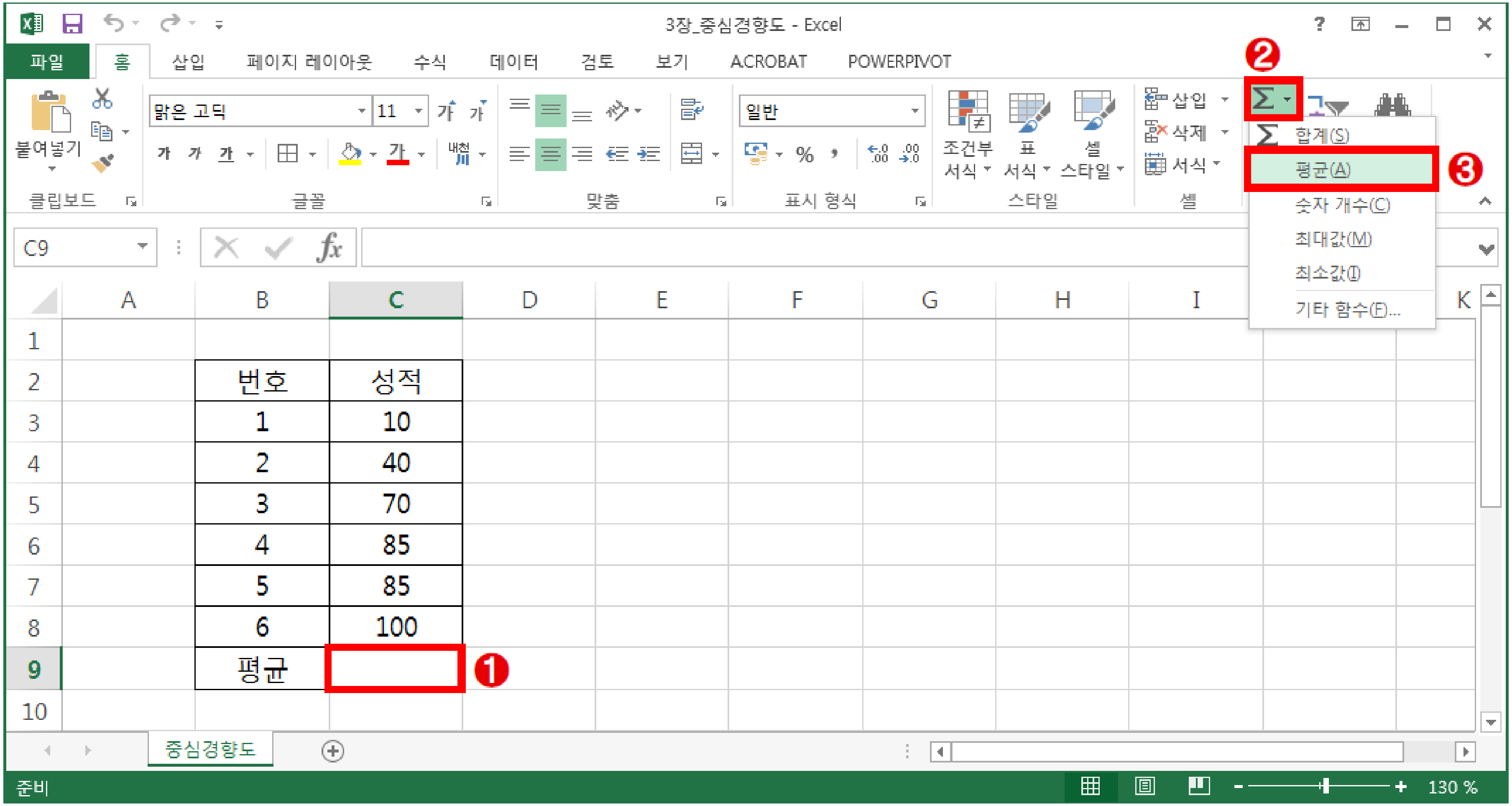
10,40,70,85,85,100 → 중간값은 70과 85

$\therefore \frac{70+85}{2}=77.5$ → A학생의 성적이 반에서 과반 이상에 들지 못했으므로 부모님으로부터 선물을 받을 수 없다.

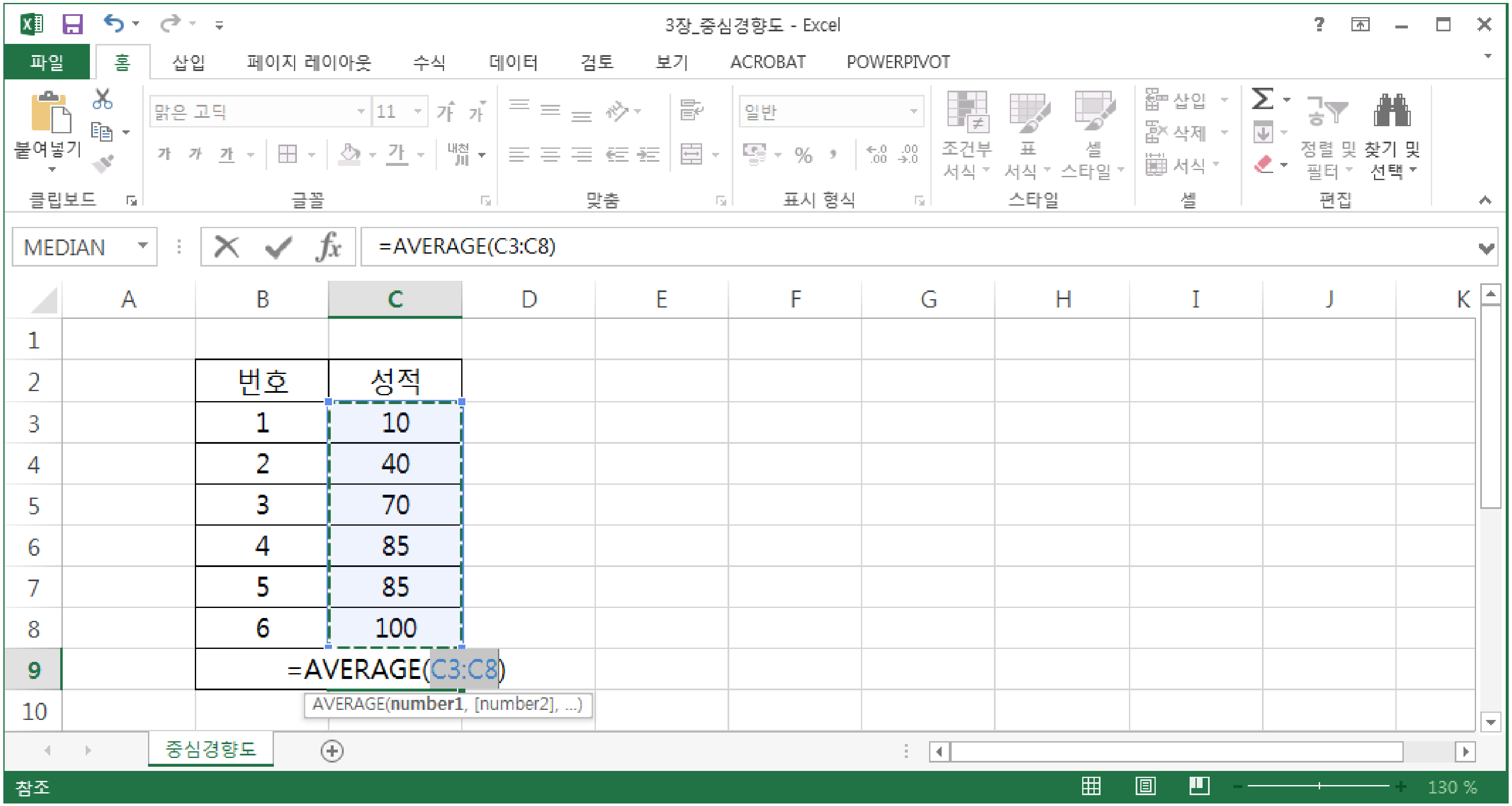
■ 최빈값

두 학생이 85점 → 85점보다 낮으므로 부모님으로부터 선물을 받을 수 없다.

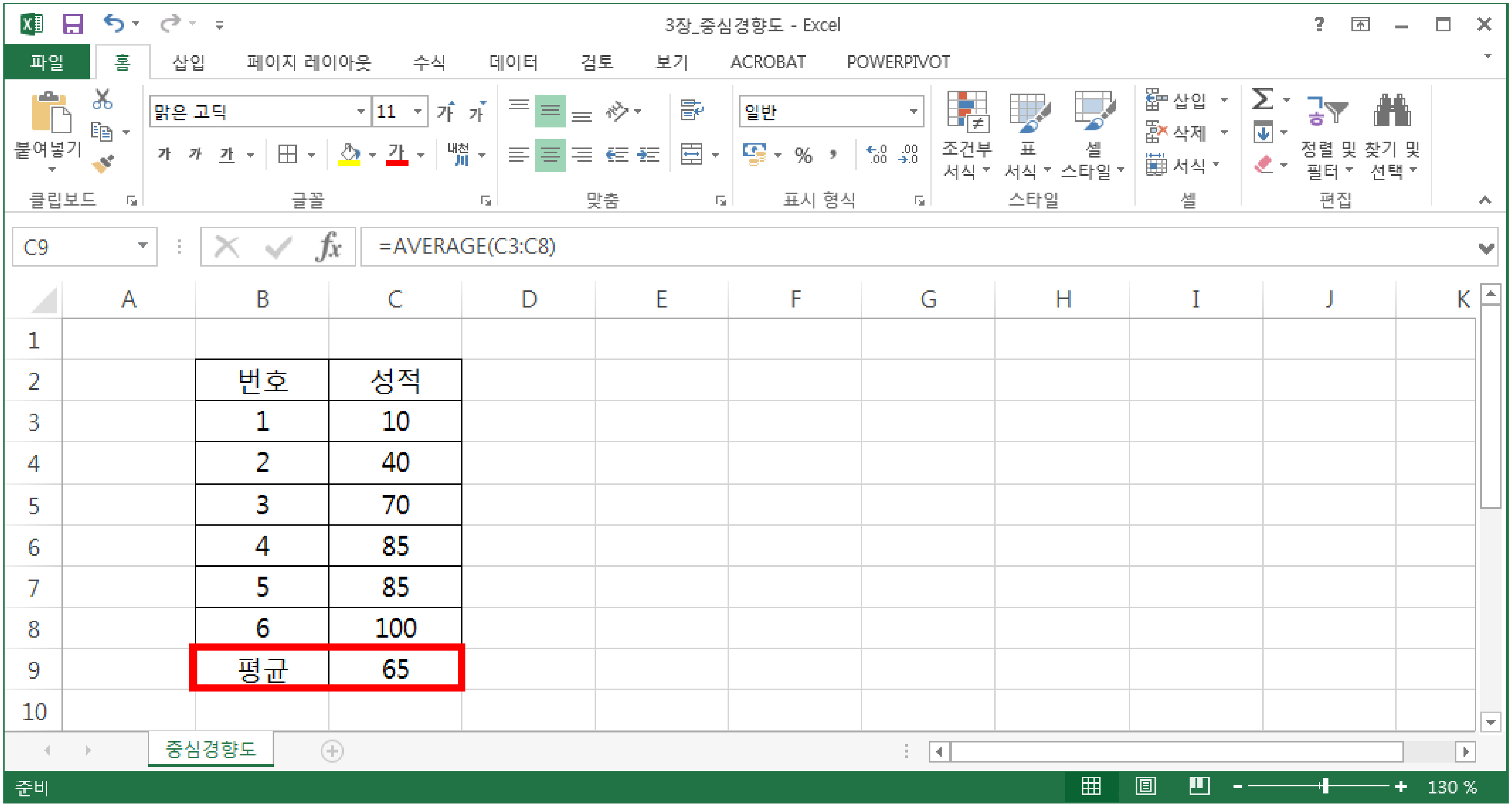
중심경향도 (Excel 풀이-평균)



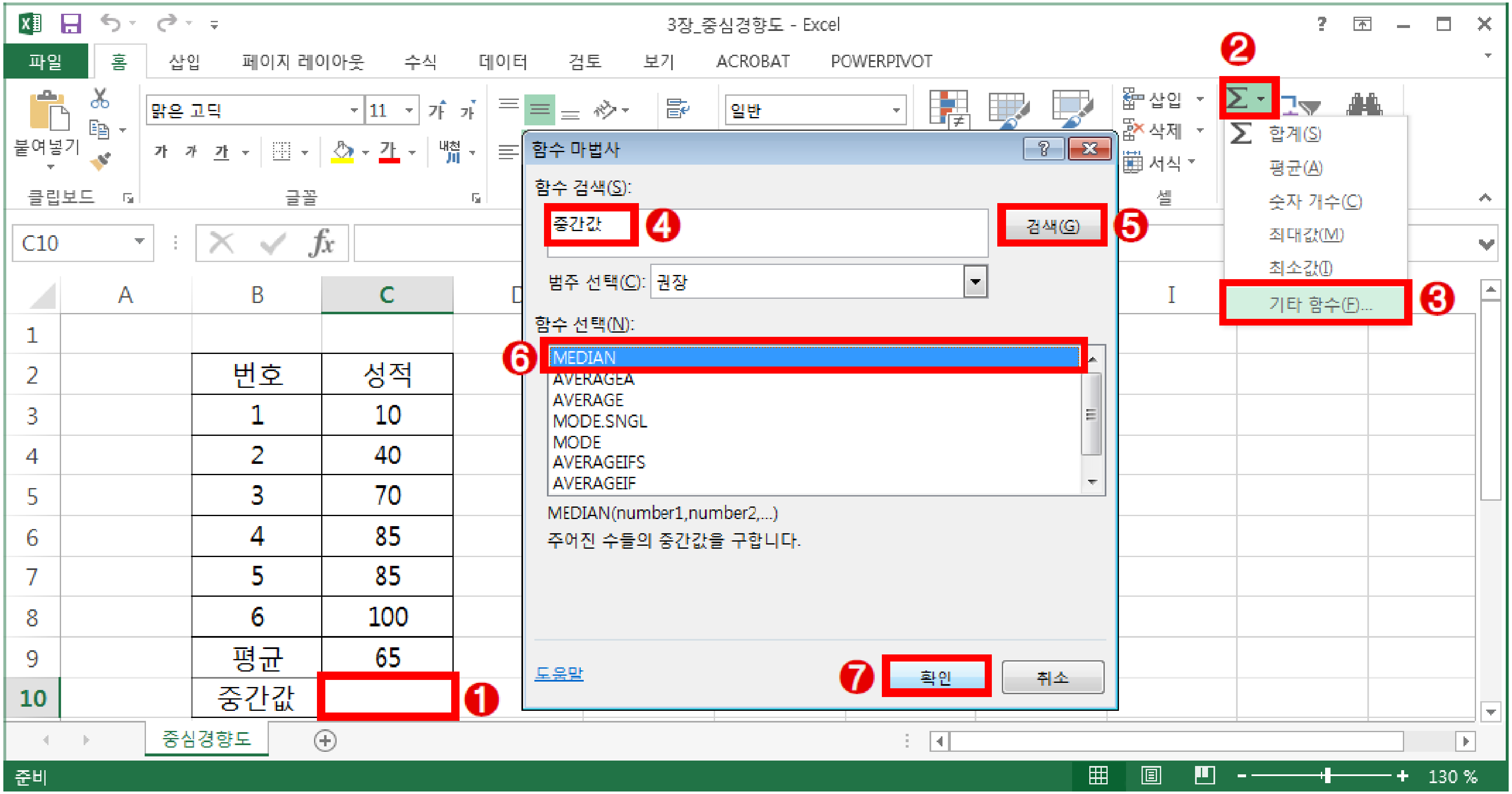
중심경향도 (Excel 풀이-평균)



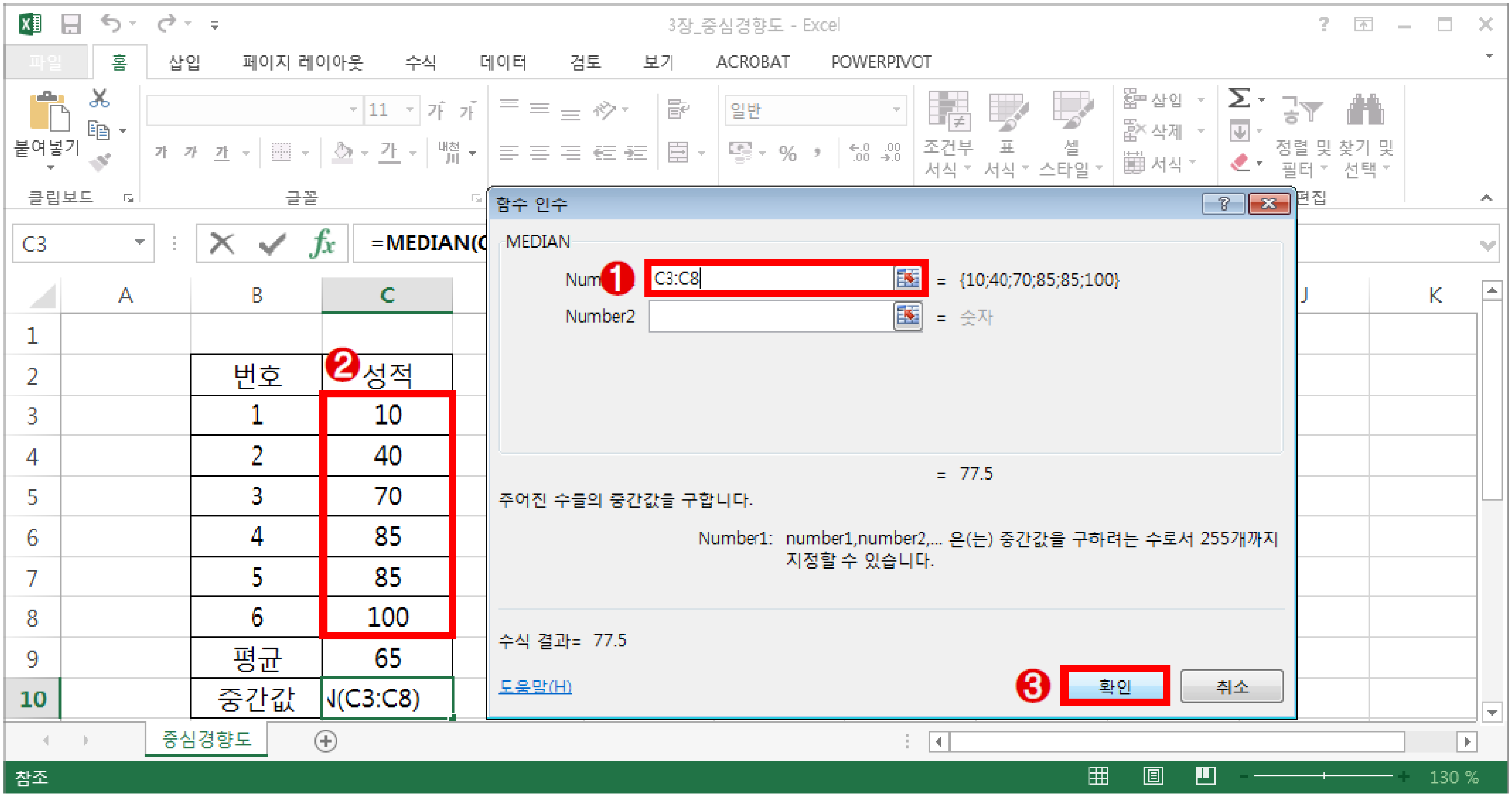
중심경향도 (Excel 풀이-평균 완성)



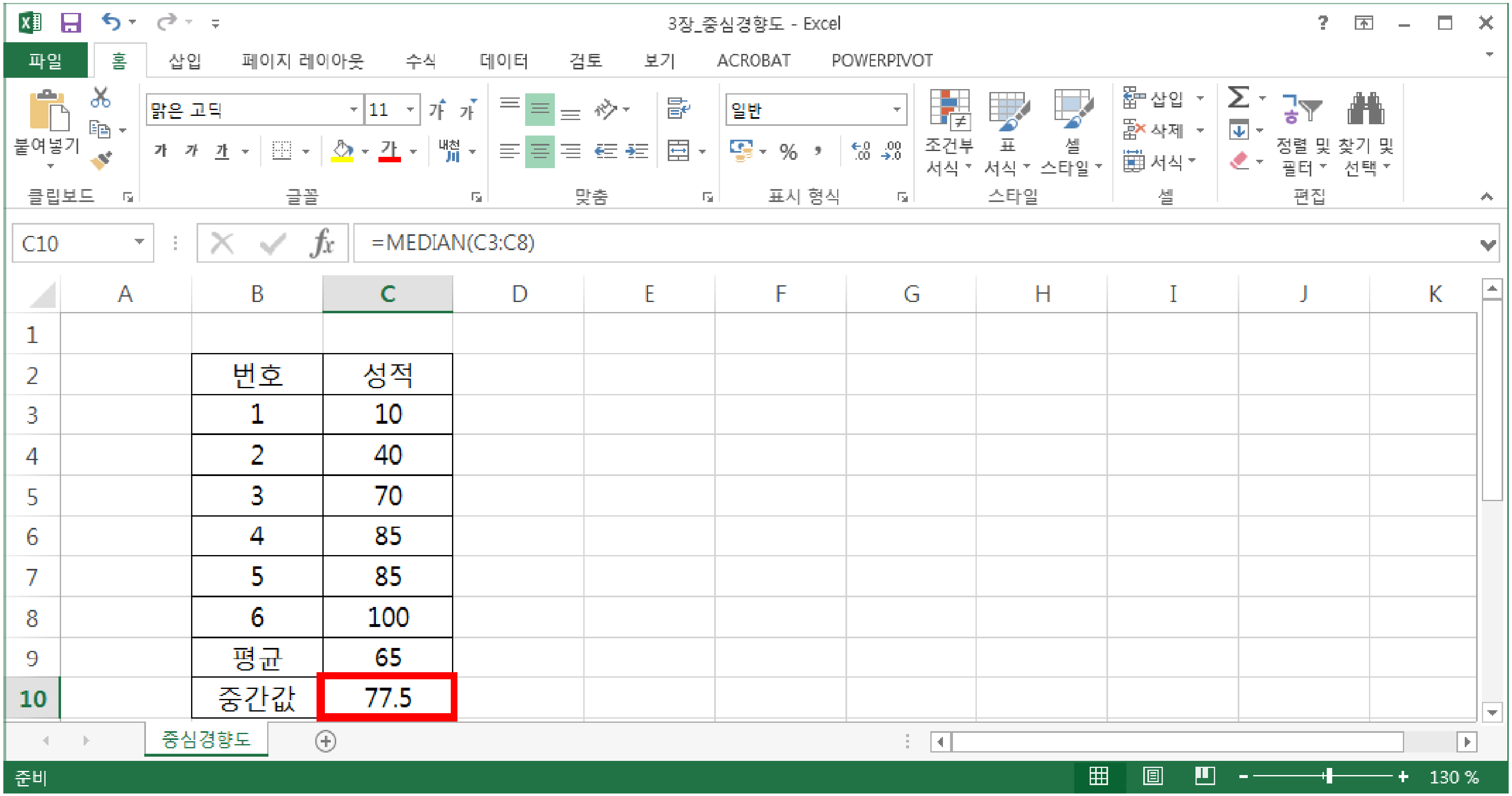
중심경향도 (Excel 풀이-중간값)



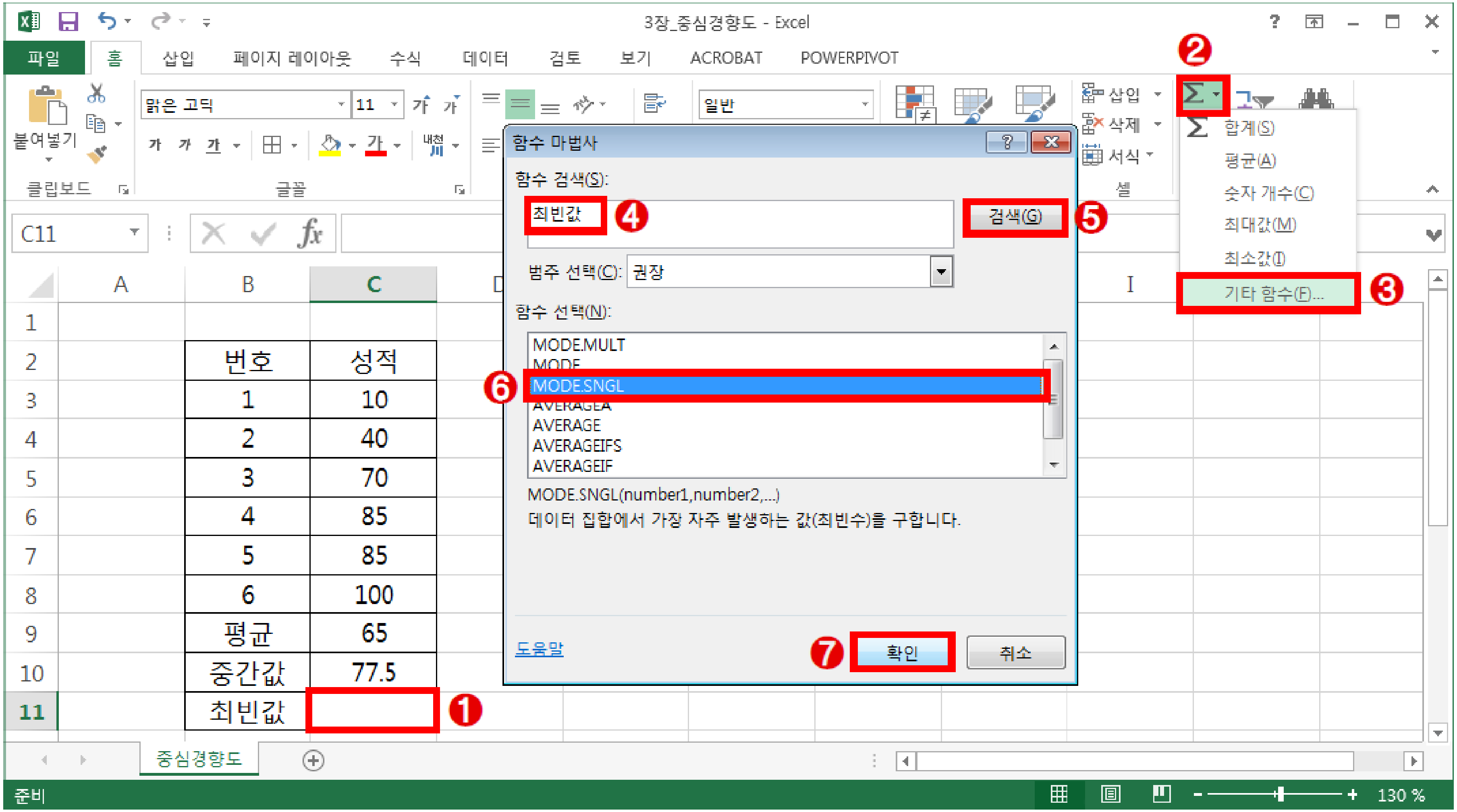
중심경향도 (Excel 풀이-중간값)



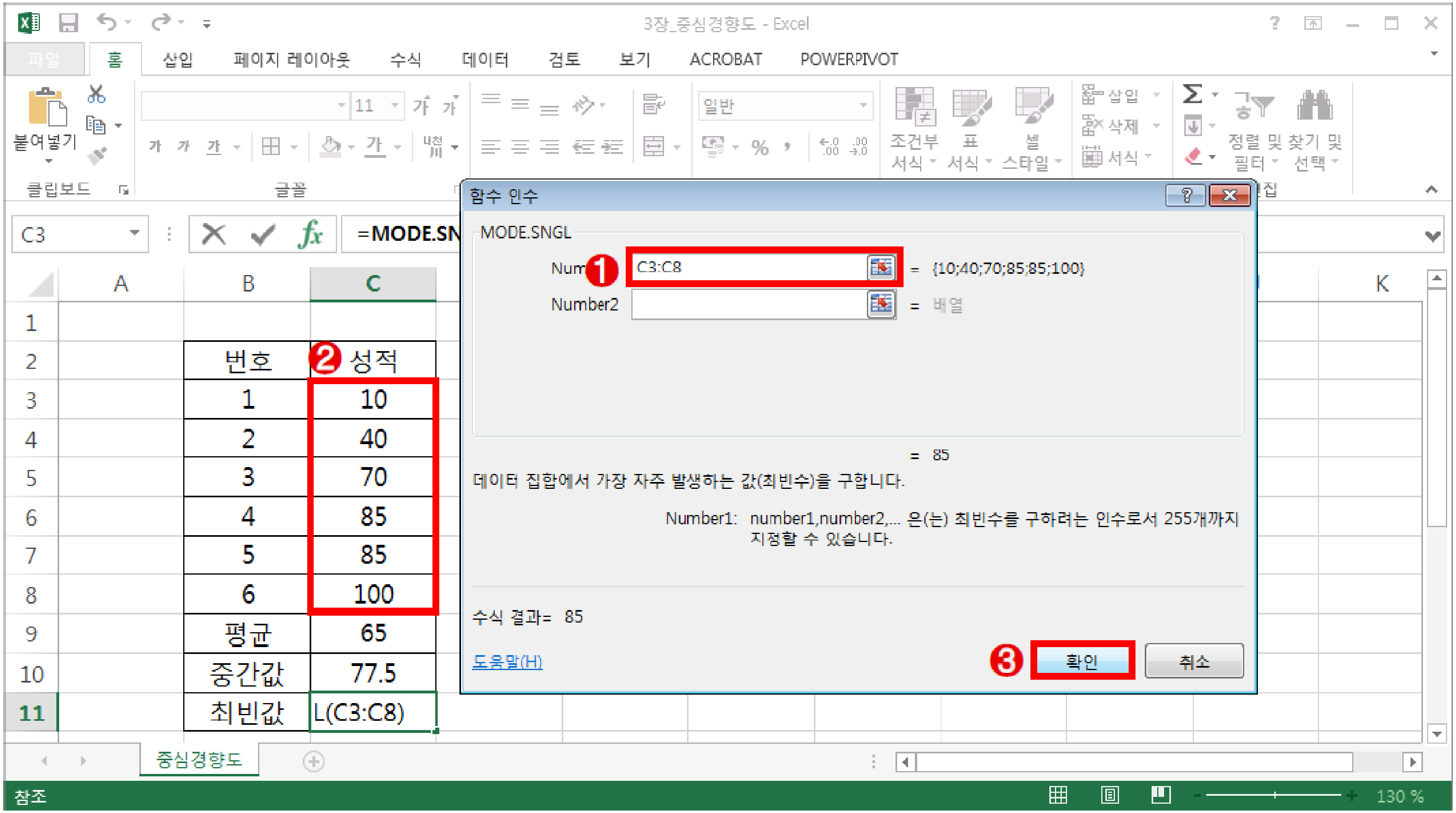
중심경향도 (Excel 풀이-중간값 완성)



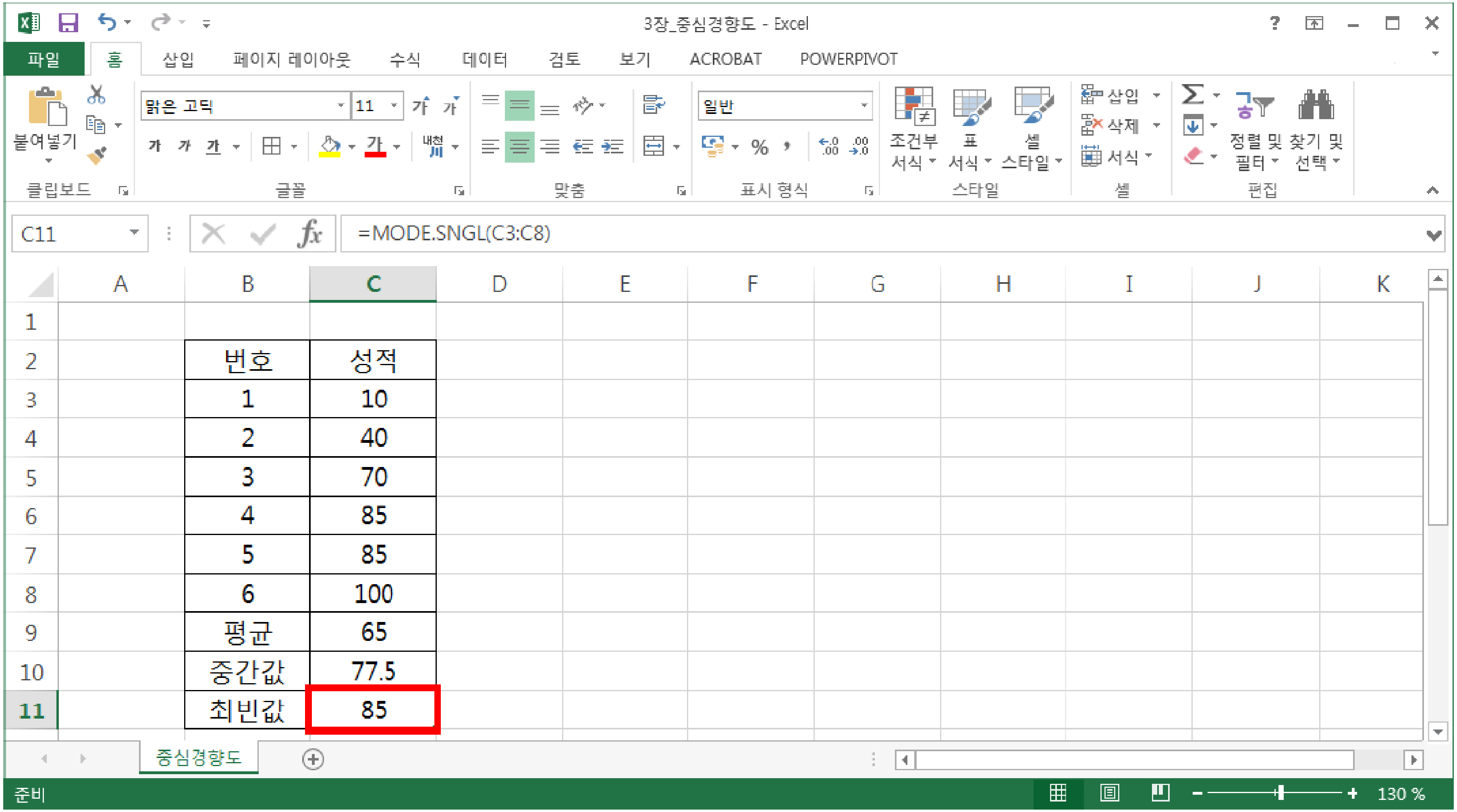
중심경향도 (Excel 풀이-최빈값)



중심경향도 (Excel 풀이-최빈값)



중심경향도 (Excel 풀이-최빈값 완성)



산포도

중심경향도만으로는 집단에 대한 성격과 분포를 파악하는데 부족하므로, 측정된 데이터가 어떻게 분포하고 있는지에 대해 파악해야 데이터를 제대로 이해하기 위하여 표본이 가지는 분포의 정도를 나타내는 산포도(dispersion)를 확인해야 한다.

→ 통계학에서 산포의 정도를 나타내는 지표에는 분산, 표준편차, 범위, 사분위수, 백분위수 등이 있음.

산포도

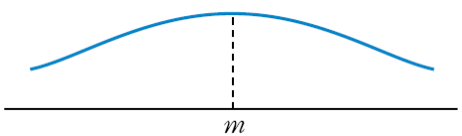
■ 모분산

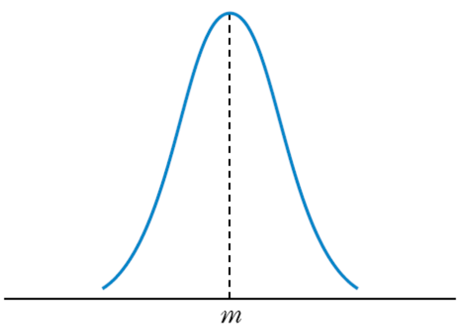
모평균과 모집단의 개별 측정치들 간의 차를 구해서 제곱하여 모두 더한 후,

그 값을 다시 모집단을 구성하는 개수로 나누어 계산

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

(집단의 개수가 n, 모평균이 μ, 모분산은 σ^2)





표본의 분포 특성을 잘 드러내지 못하는 평균의 단점을 해소하기 위해서는
평균과 각 표본들이 얼마나 떨어져 있는지를 측정한 차이(편차)를 확인해야 한다.

산포도

예제 3-4 **분산 계산** 준비파일 | 3장_신장.xlsx

다음은 A반과 B반의 신장을 조사한 자료이다. 이때 A반과 B반의 평균과 분산을 구하고, 그 특성과 차이에 대해 설명하라.

단위 : cm

반	1	2	3	4	5	6	7	8	9	10
A	168	160	170	162	168	163	164	167	161	166
B	175	179	152	153	173	158	175	154	172	157

산포도 (모분산 풀이)

■ 평균

$$\bar{x}_A = \frac{168 + 160 + 170 + 162 + 168 + 163 + 164 + 167 + 161 + 166}{10} = 164.9$$

$$\bar{x}_B = \frac{175 + 179 + 152 + 153 + 173 + 158 + 175 + 154 + 172 + 157}{10} = 164.8$$

→ 평균 신장이 164.9cm와 164.8cm로 0.1cm(1mm) 정도의 미세한 차이

산포도 (모분산 풀이)

■ 분산

$$\sigma_A^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - 164.9)^2$$

$$= \frac{(168-164.9)^2 + (160-164.9)^2 + \cdots + (166-164.9)^2}{10} = 10.29$$

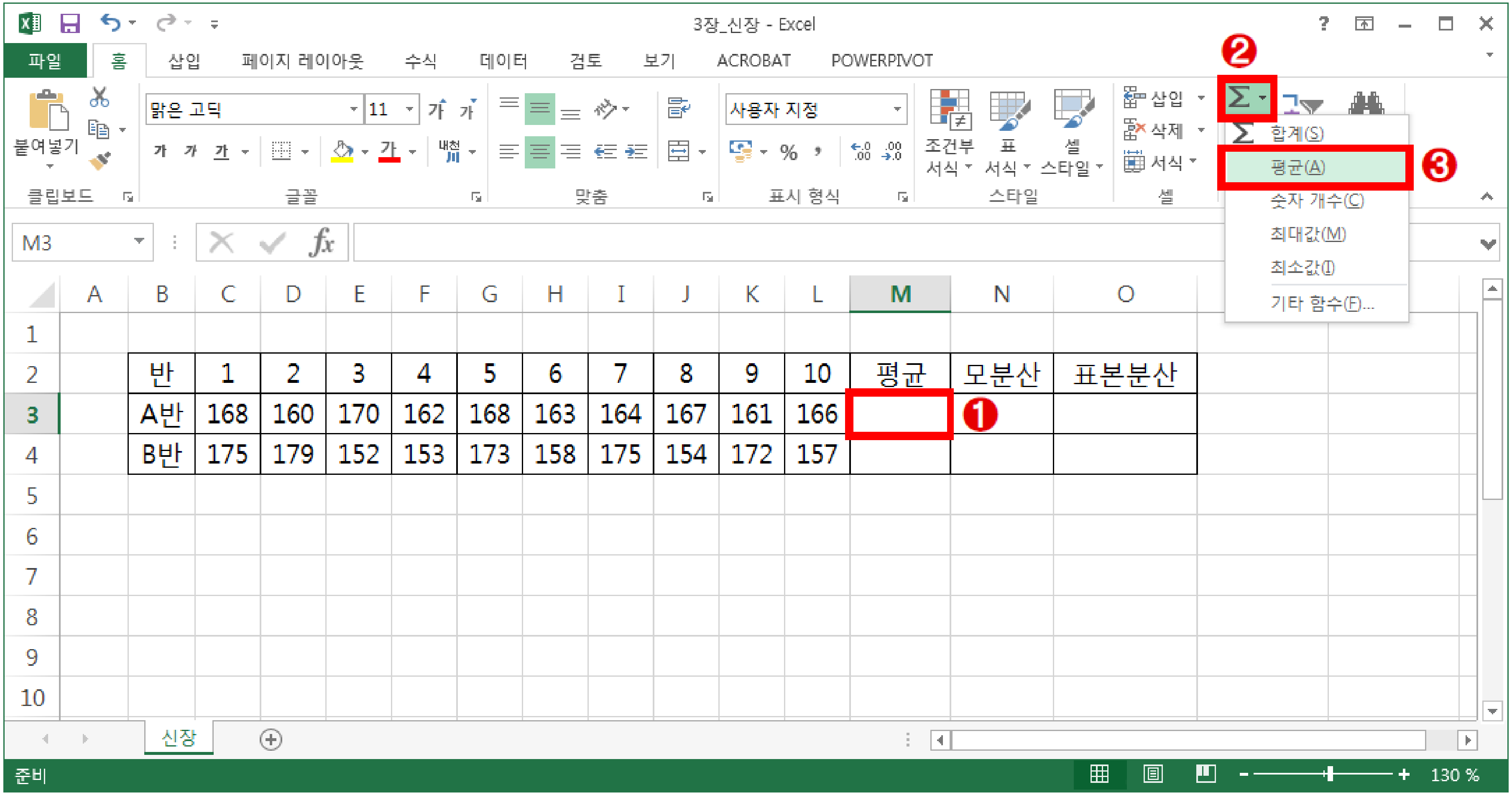
$$\sigma_B^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - 164.8)^2$$

$$= \frac{(175-164.8)^2 + (179-164.8)^2 + \cdots + (157-164.8)^2}{10} = 105.56$$

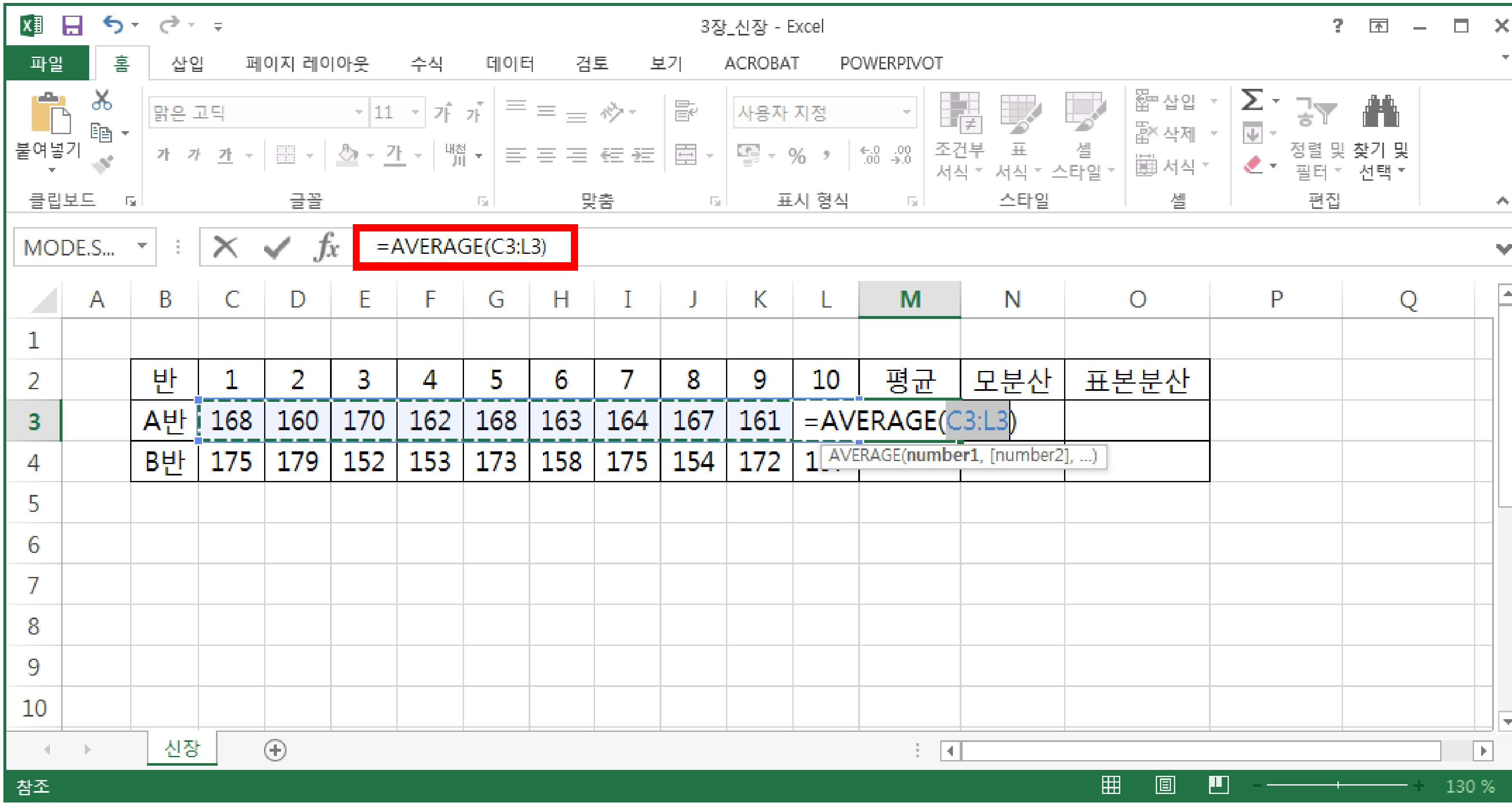
→ A반에서는 거의 모든 학생의 키가 비슷하지만,

B반에서는 키가 큰 학생과 작은 학생의 차이가 크게 난다는 것을 의미

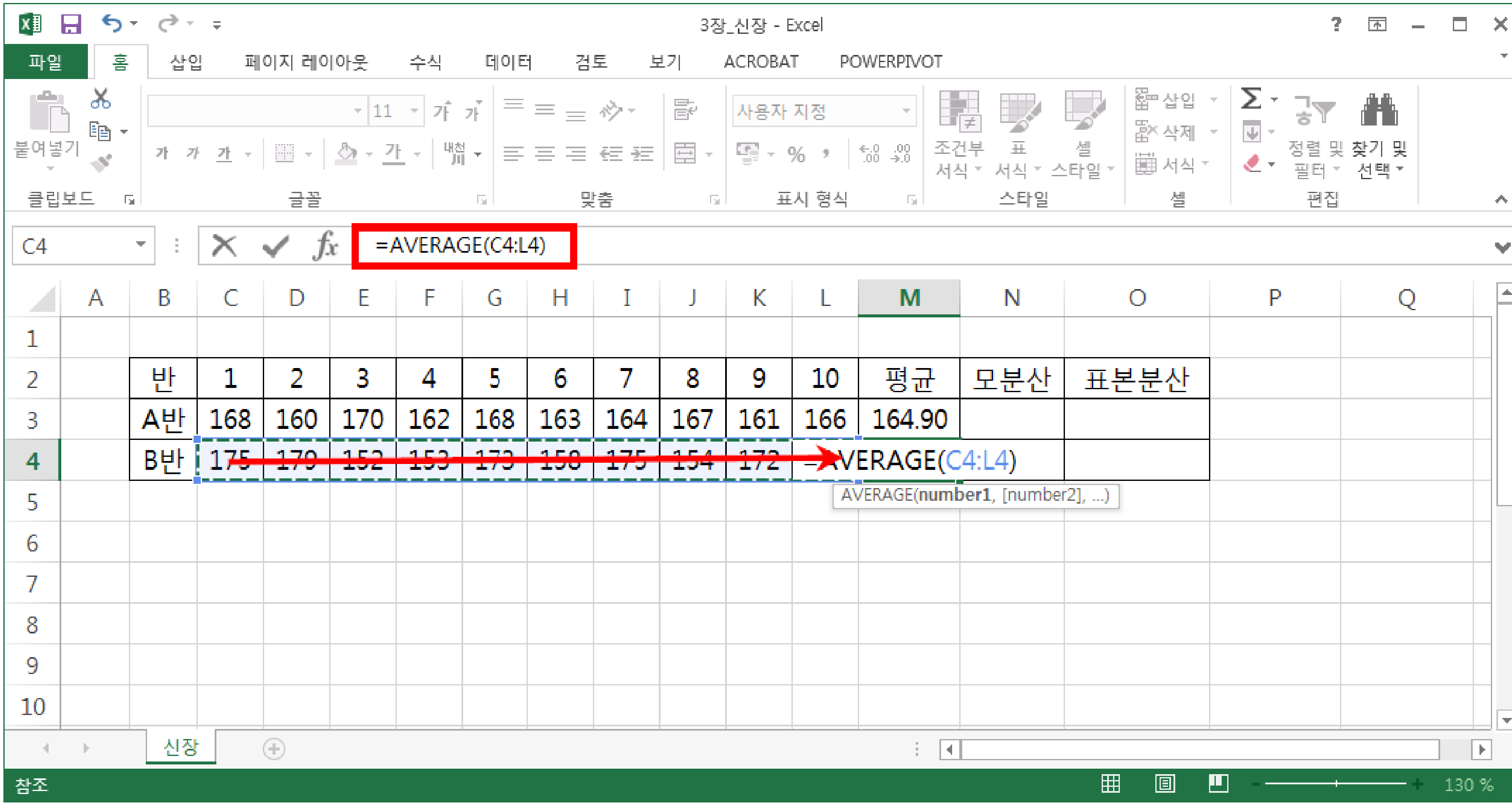
산포도 (모분산 Excel 풀이)



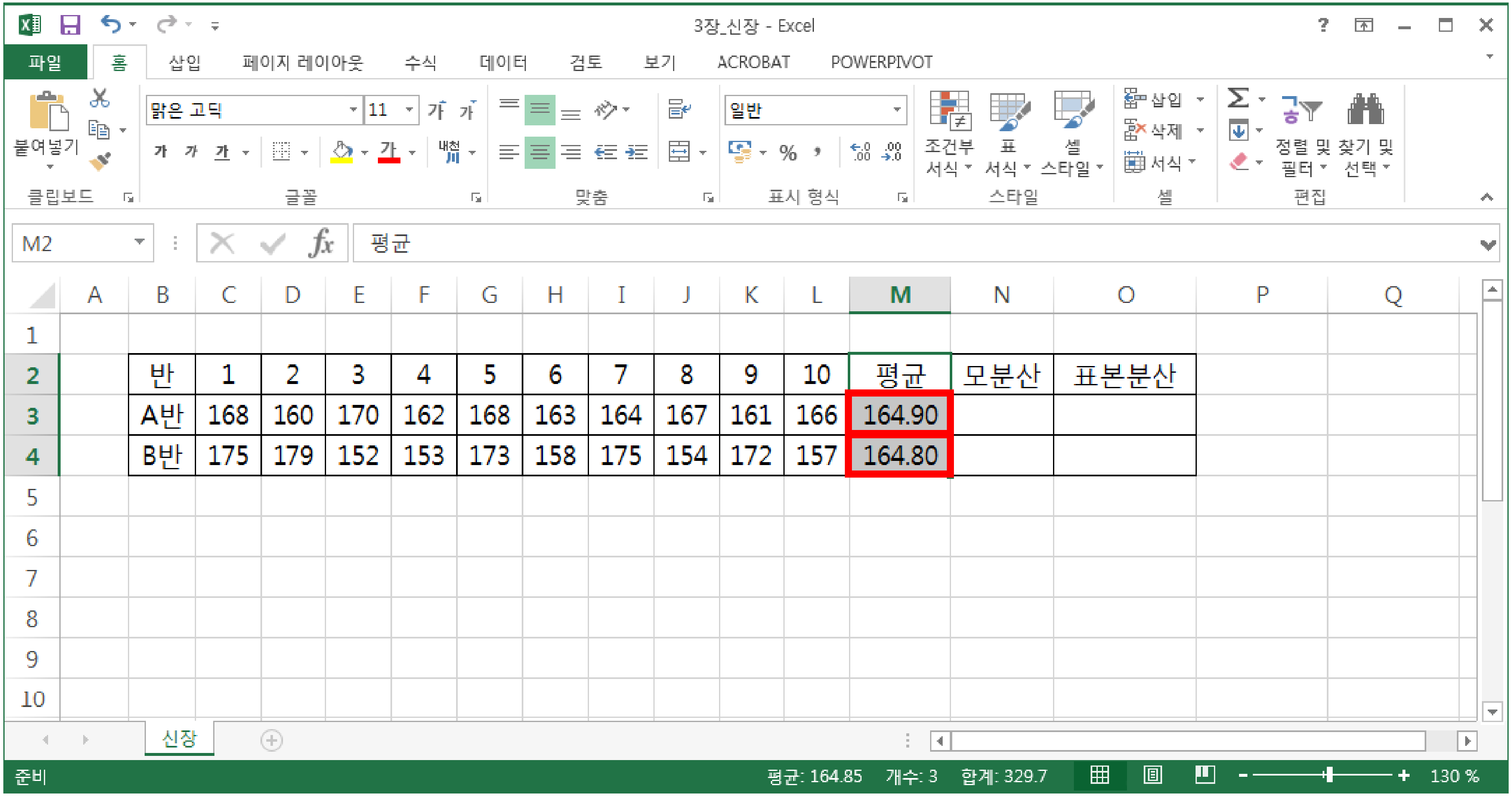
산포도 (모분산 Excel 풀이)



산포도 (모분산 Excel 풀이)

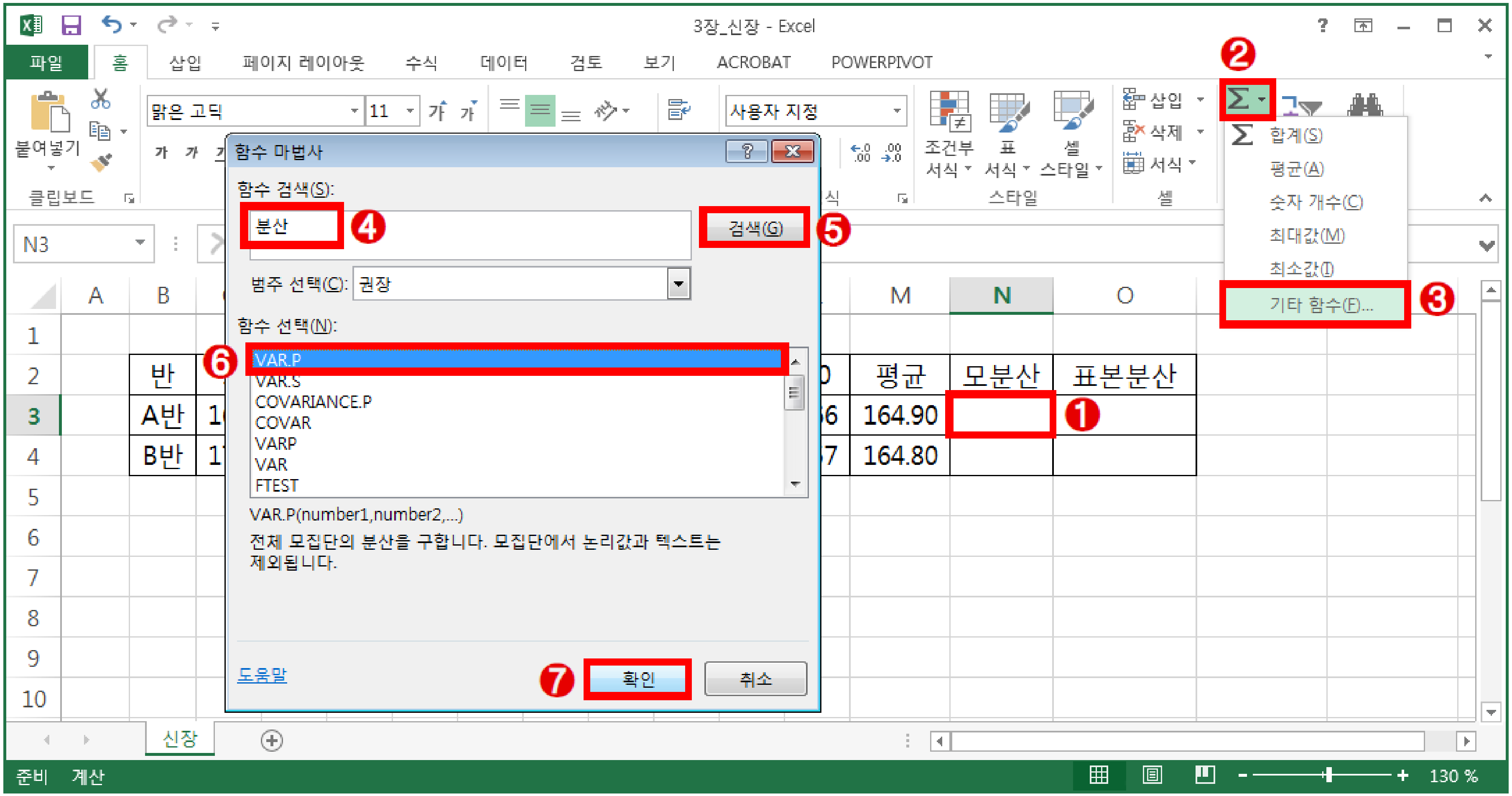


산포도 (모분산 Excel 풀이)

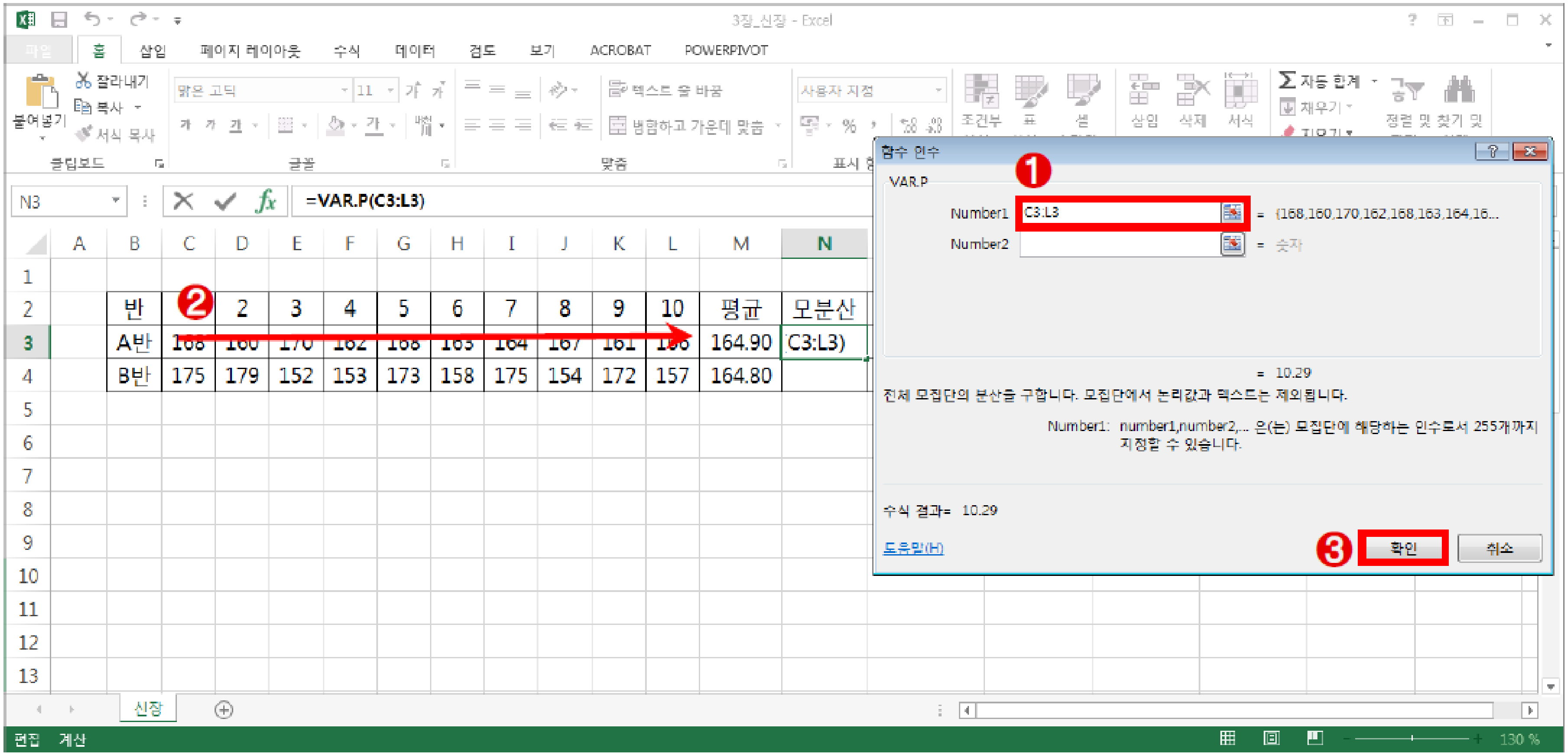


반	1	2	3	4	5	6	7	8	9	10	평균	모분산	표본분산
A반	168	160	170	162	168	163	164	167	161	166	164.90		
B반	175	179	152	153	173	158	175	154	172	157	164.80		

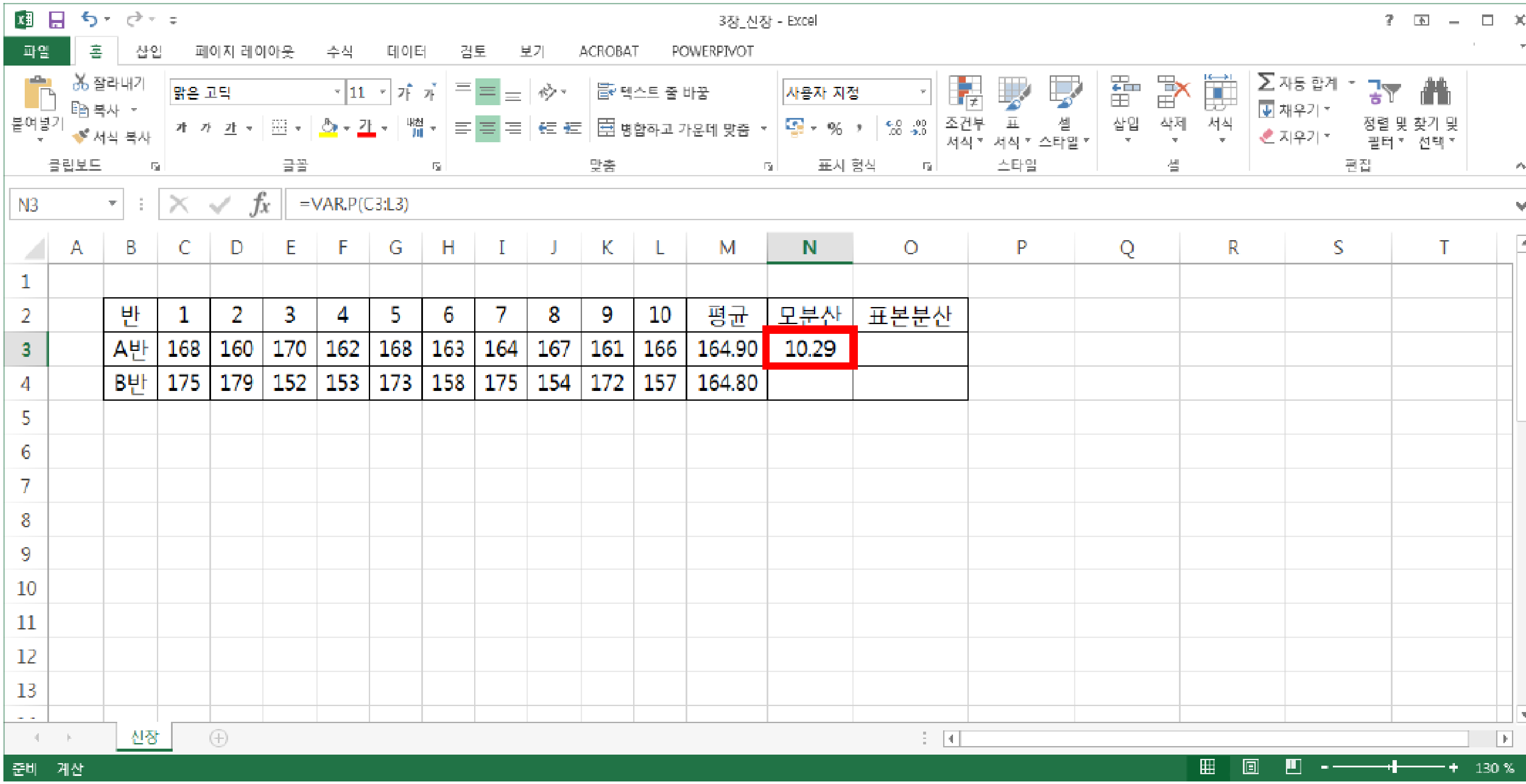
산포도 (모분산 Excel 풀이)



산포도 (모분산 Excel 풀이)



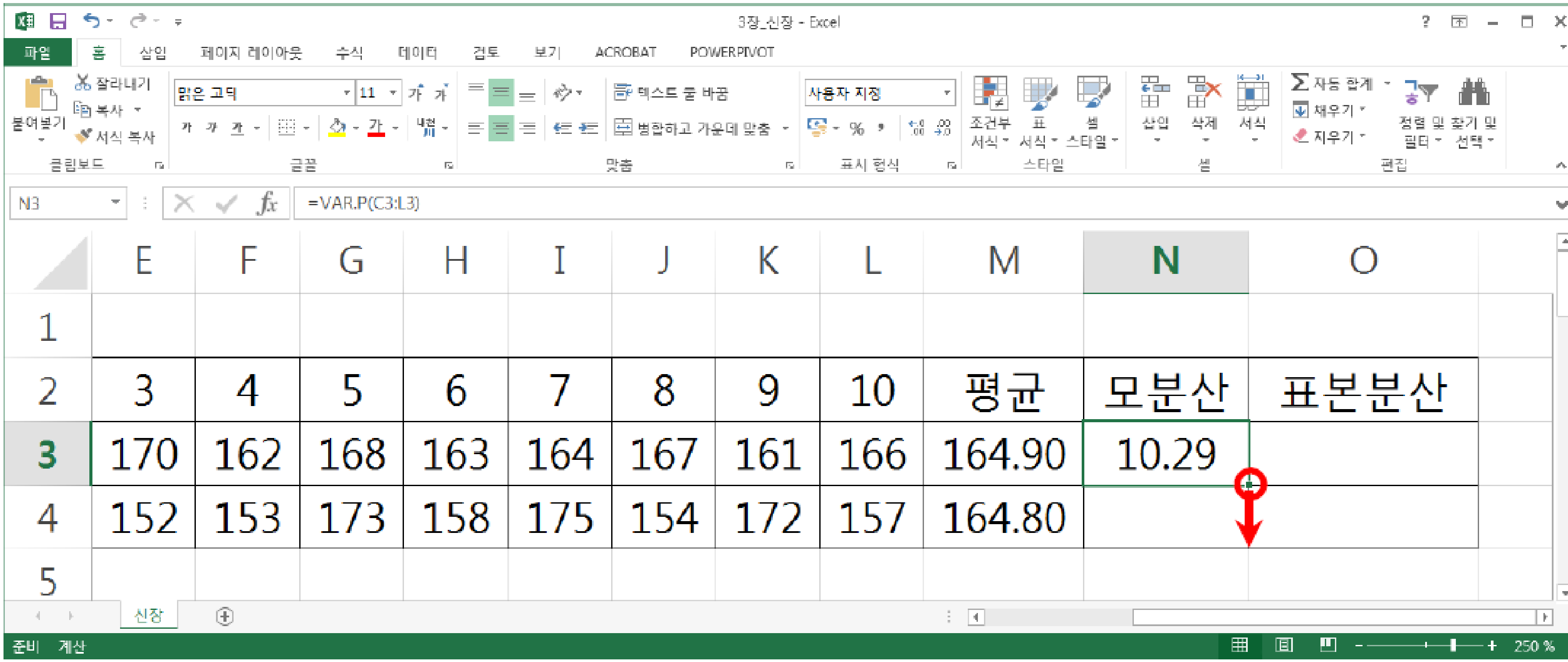
산포도 (모분산 Excel 풀이)



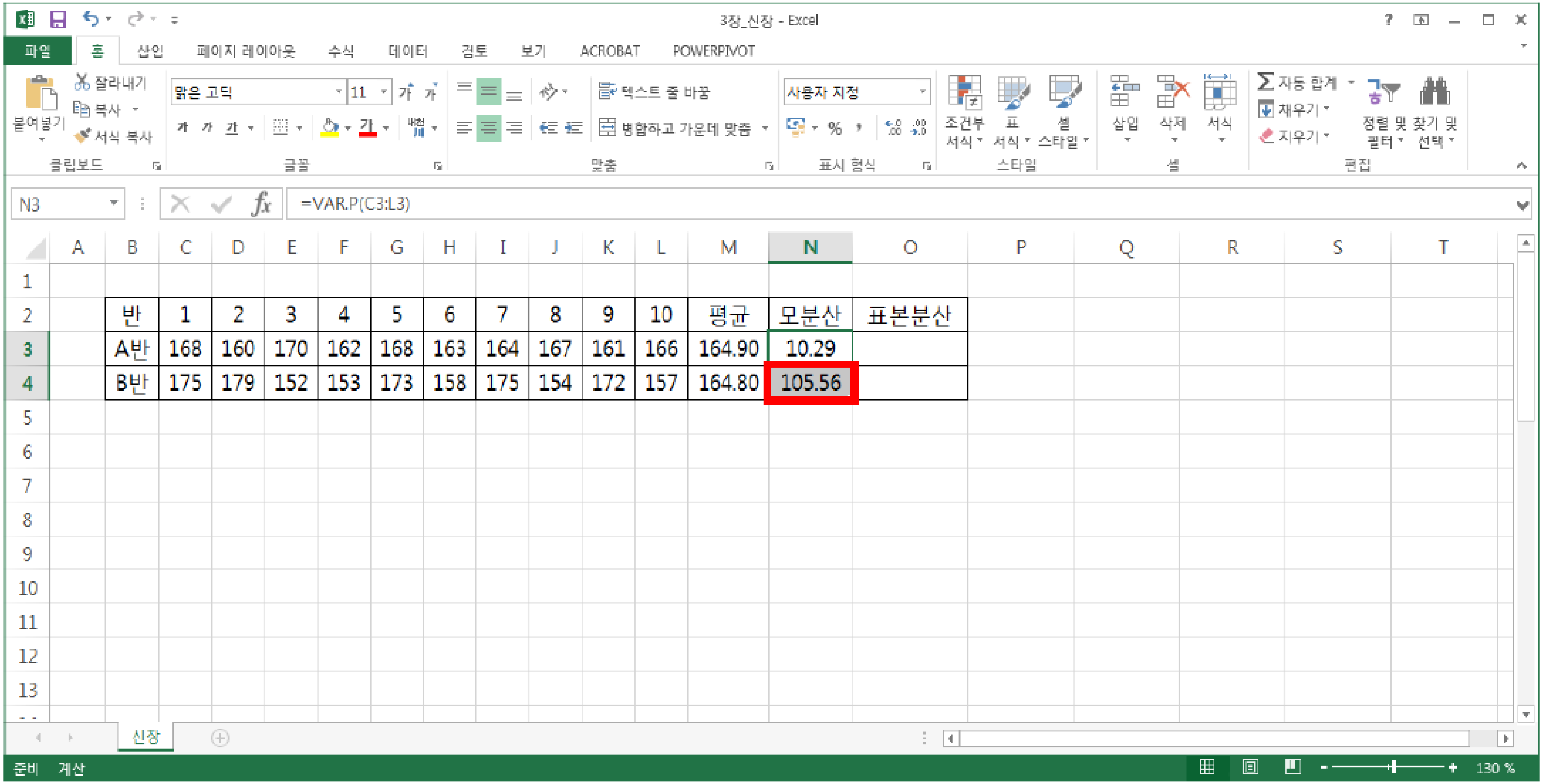
N3 =VAR.P(C3:L3)

반	1	2	3	4	5	6	7	8	9	10	평균	모분산	표본분산
A반	168	160	170	162	168	163	164	167	161	166	164.90	10.29	
B반	175	179	152	153	173	158	175	154	172	157	164.80		

산포도 (모분산 Excel 풀이)



산포도 (모분산 Excel 풀이 완성)



반	1	2	3	4	5	6	7	8	9	10	평균	모분산	표본분산
A반	168	160	170	162	168	163	164	167	161	166	164.90	10.29	
B반	175	179	152	153	173	158	175	154	172	157	164.80	105.56	

산포도

■ 표본분산

모집단을 기준으로 하지 않고, 표본을 선정해서 표본의 개수 (n-1)로 계산한 분산을 표본분산(sample variance)이라 한다.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad (\overline{x} : \text{평균})$$

예제 3-5 표본분산 계산

준비파일 | 3장_신장.xlsx

[예제 3-4]에 주어진 조건을 이용하여 A반과 B반의 표본분산을 구하고, 그 특성과 차이에 대해 설명하라.

산포도 (표본분산 풀이)

■ 표본분산

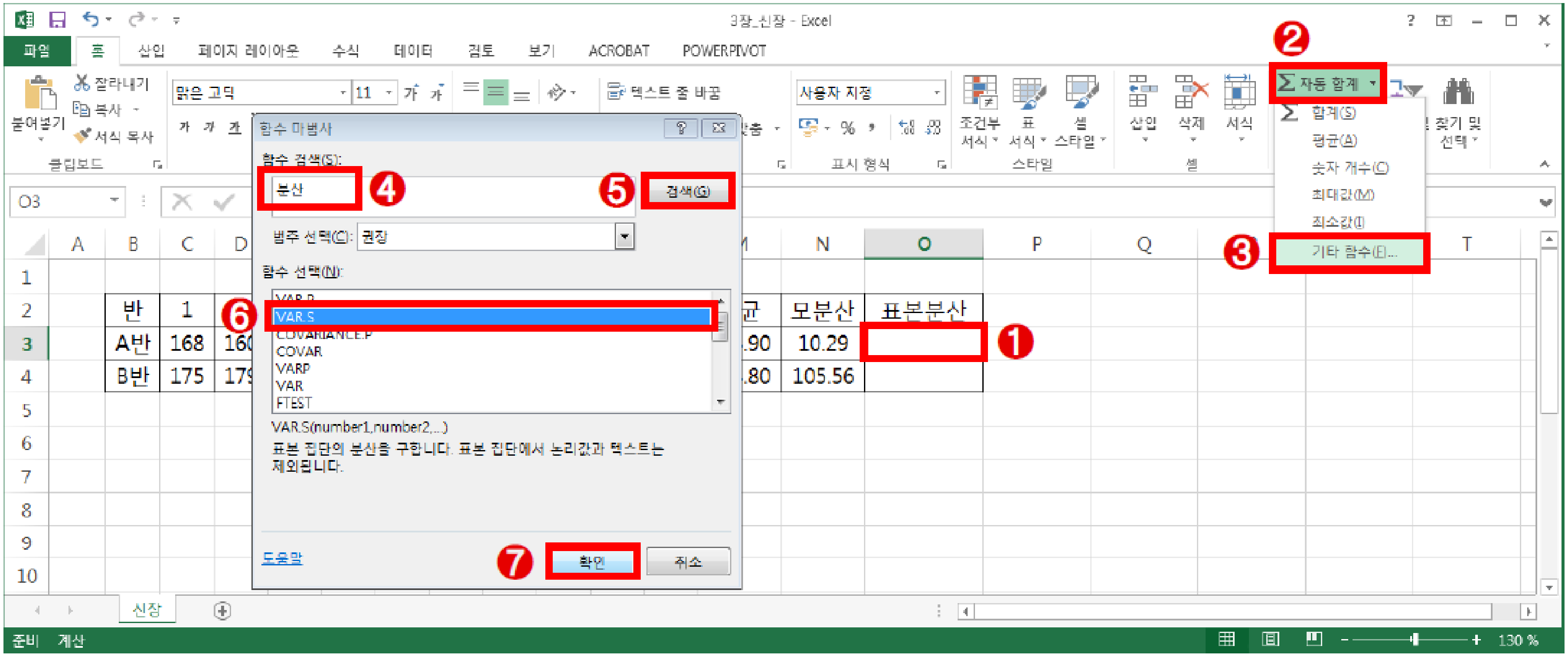
$$s_A^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - 164.9)^2$$

$$= \frac{(168-164.9)^2 + (160-164.9)^2 + \cdots + (166-164.9)^2}{9} = 11.433$$

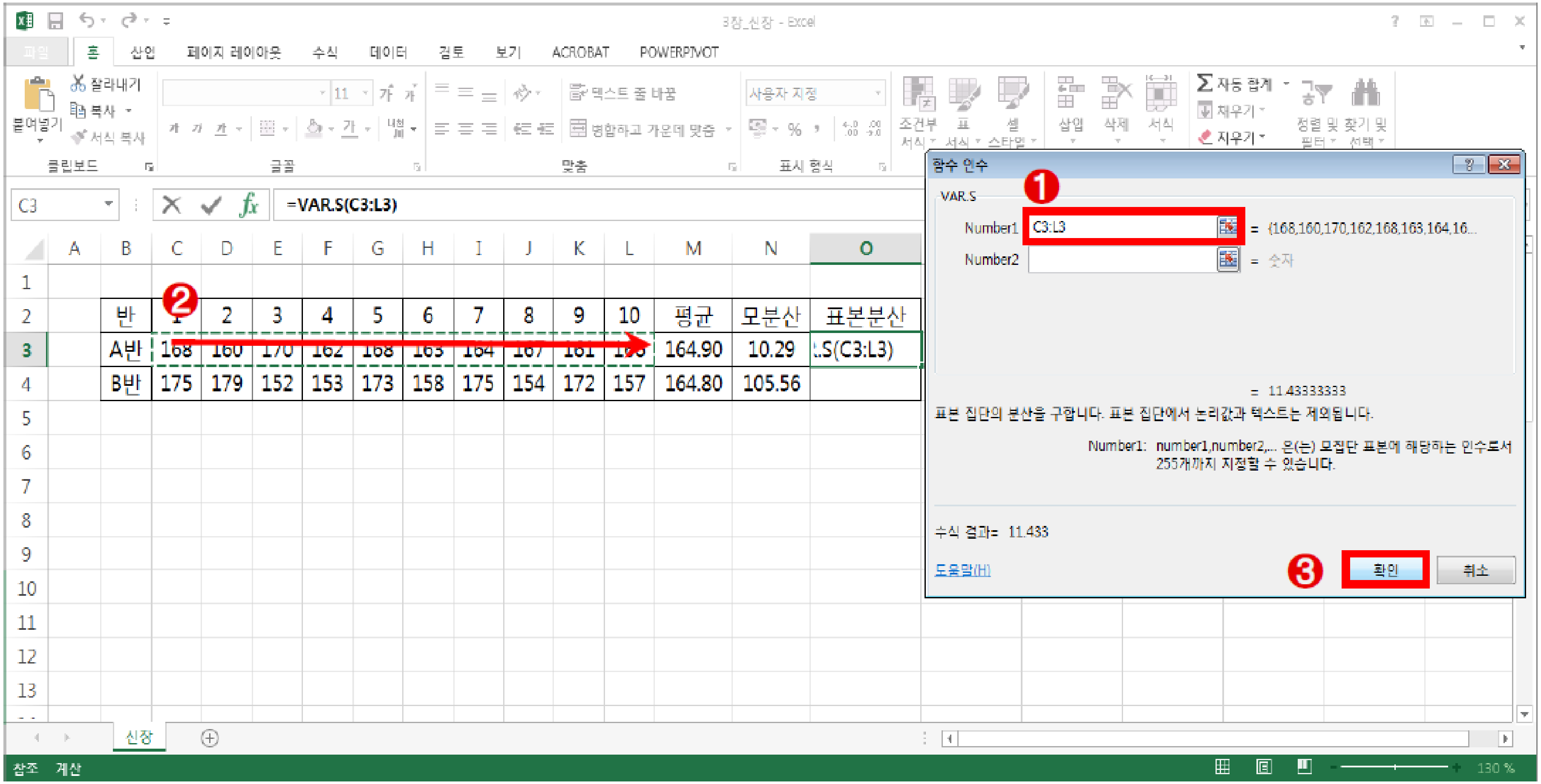
$$s_B^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - 164.8)^2$$

$$= \frac{(175-164.8)^2 + (179-164.8)^2 + \cdots + (157-164.8)^2}{9} = 117.289$$

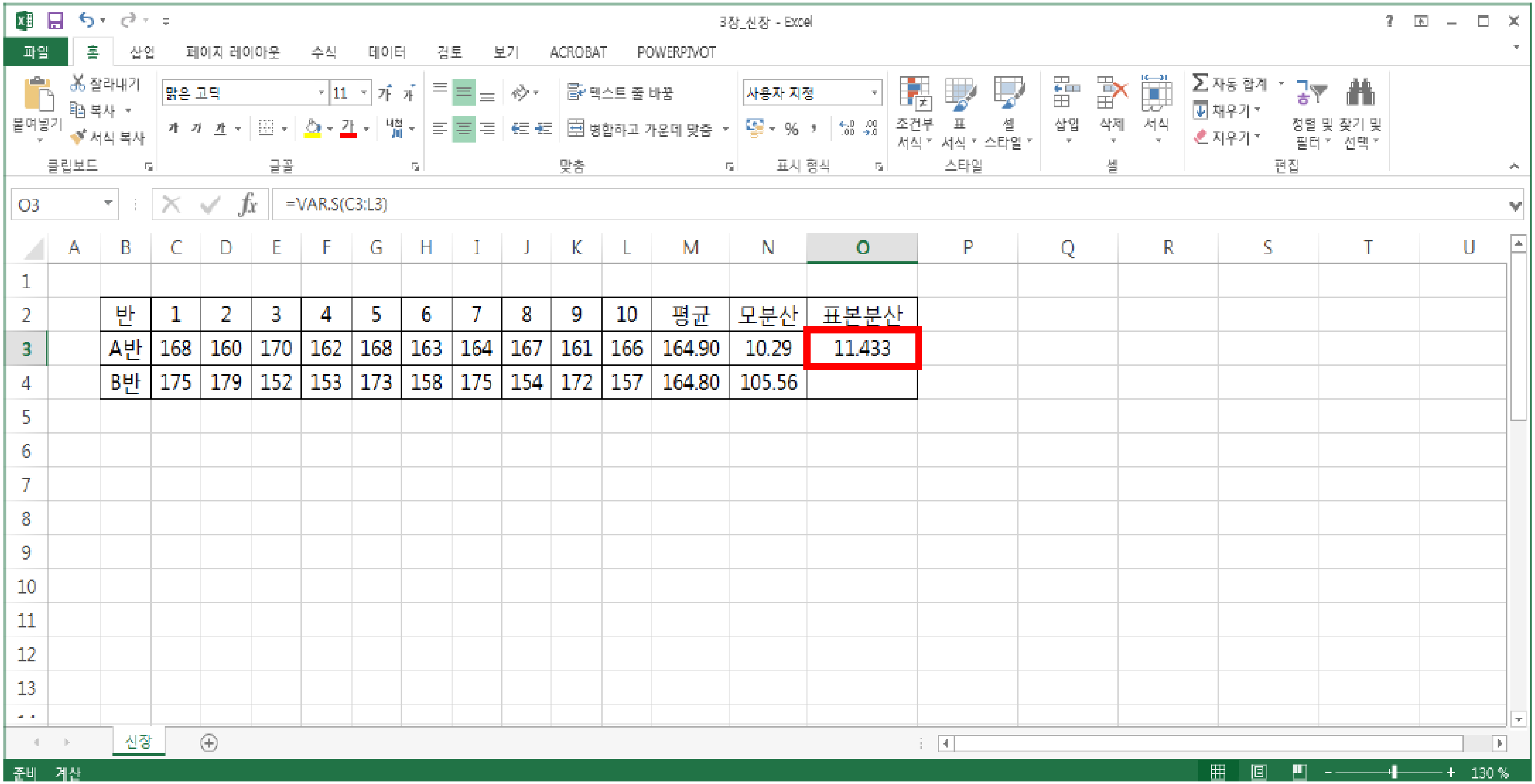
산포도 (표본분산 Excel 풀이)



산포도 (표본분산 Excel 풀이)



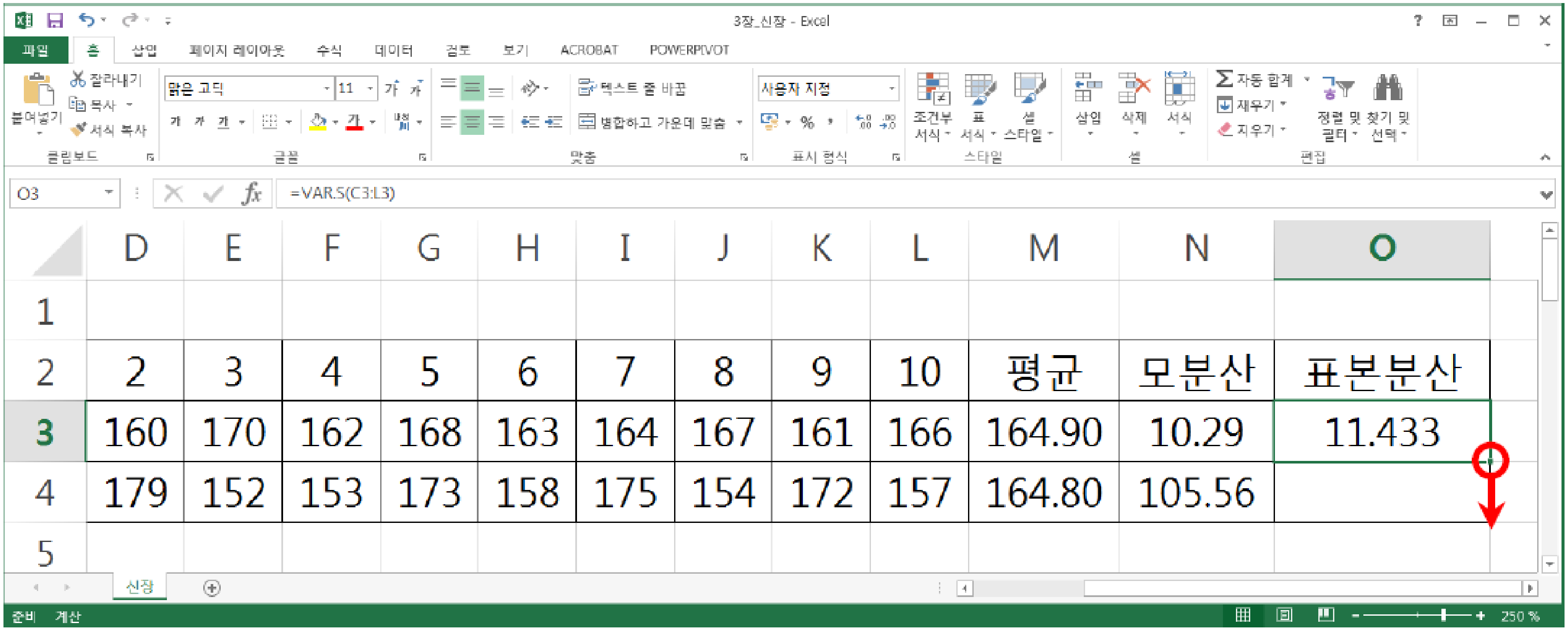
산포도 (표본분산 Excel 풀이)



O3 =VAR.S(C3:L3)

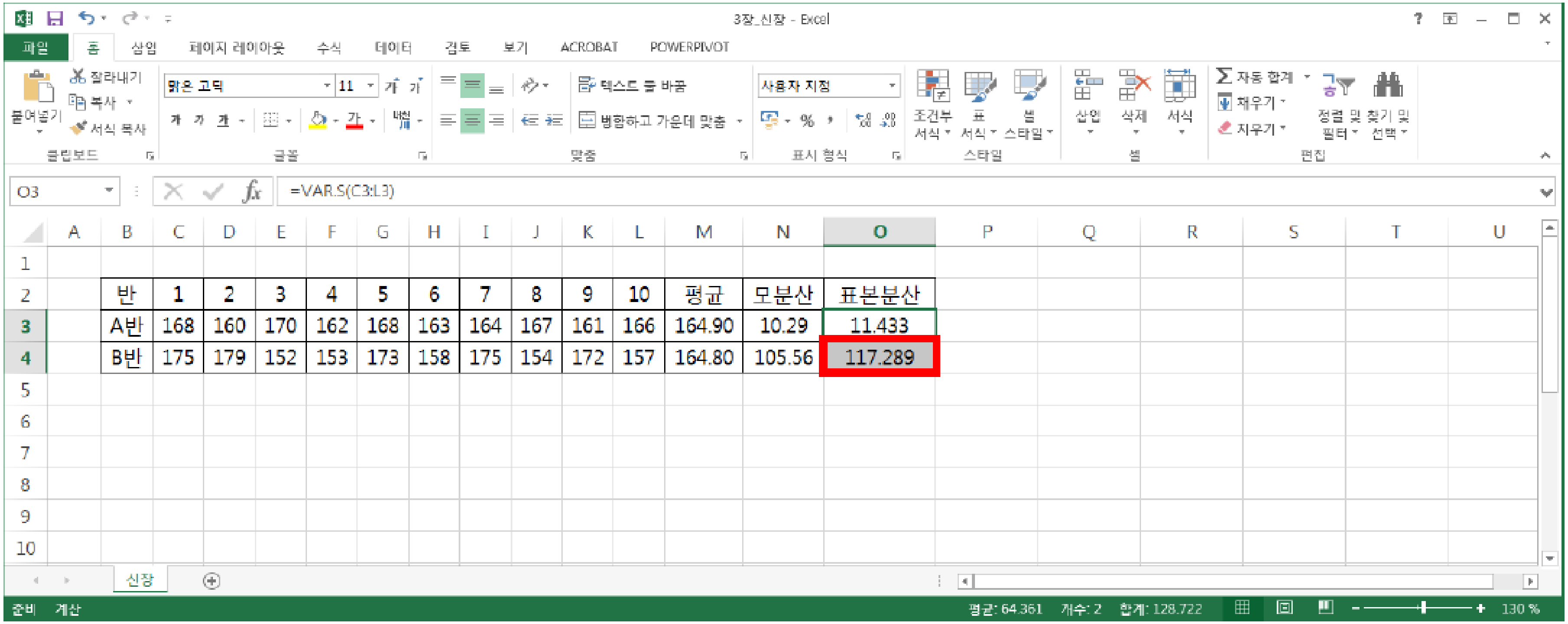
반	1	2	3	4	5	6	7	8	9	10	평균	모분산	표본분산
A반	168	160	170	162	168	163	164	167	161	166	164.90	10.29	11.433
B반	175	179	152	153	173	158	175	154	172	157	164.80	105.56	

산포도 (표본분산 Excel 풀이)



	D	E	F	G	H	I	J	K	L	M	N	O
1												
2	2	3	4	5	6	7	8	9	10	평균	모분산	표본분산
3	160	170	162	168	163	164	167	161	166	164.90	10.29	11.433
4	179	152	153	173	158	175	154	172	157	164.80	105.56	
5												

산포도 (표본분산 Excel 풀이 완성)



O3 =VAR.S(C3:L3)

반	1	2	3	4	5	6	7	8	9	10	평균	모분산	표본분산
A반	168	160	170	162	168	163	164	167	161	166	164.90	10.29	11.433
B반	175	179	152	153	173	158	175	154	172	157	164.80	105.56	117.289

산포도

■ 표준편차

분산과 편차의 개념은 평균으로부터 측정치들이 어느 정도 흩어져 있는지의 정도를 나타내는 것이다. 다만 편차는 평균을 기준으로 음(-)과 양(+)으로 흩어져서 총합이 0이 되니, 이를 피하기 위해 편차에 제곱을 하는 것

→분산값에 루트(√)를 씌워 제곱근을 만들면 표준편차(standard deviation)를 계산