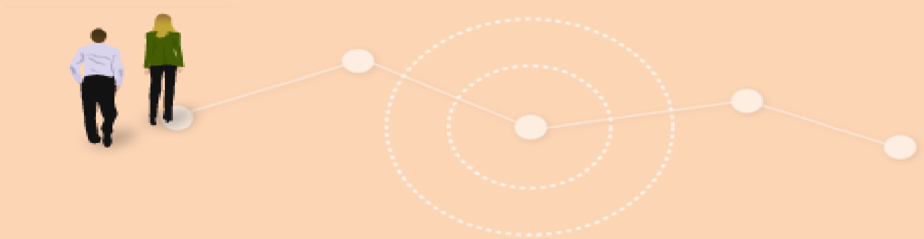
Chapter 06

추정 Estimation

목 차

01 점추정과 구간추정

:: Keywords 점추정 | 구간추정 | 신뢰구간



추정이란

추정은 정확하지는 않지만 이 정도면 될 것이라는 정도를 가늠하는 것이다. 그 정도를 가늠하는 방법은 정확하게 어떠한 수치로 나타낼 수도 있고, 시작점과 종료점으로 표현할 수도 있다.

■ 점추정(point estimation)

점추정은 모수를 특정한 수치로 표현하는 것

Ex. 통학 시간에 대해 점추정은 30분, 40분과 같이 특정한 수치로 표현

■ 구간추정(interval estimation)

구간추정은 모수를 최소값과 최대값의 범위로 추정하는 것

Ex. 통학 시간에 대해 구간추정은 30분~40분과 같이 범위로 표현

추정이란

■ 추정치(estimate)

모수를 추정하기 위해 선택된 표본을 대상으로 구체적으로 도출된 통계량

■ 추정량(estimator)

표본에서 관찰된 값으로 추정치를 계산하기 위한 도출 함수

점추정과 바람직한 점추정량의 조건

■ 점추정

가정 : 토익 시험에서 750점, 850점, 800점을 받았다.

→ 다음에 받을 성적을 추정하면 몇 점이 될까?
모든 조건이 동일하다고 했으니 3번의 시험 성적 평균인 800점으로 추정하면 무난함.
하지만 800점은 맞다고 할 수도 없고, 틀렸다고 할 수도 없다.

실제로 800점을 얻는다면 정확한 추정
그 외의 점수를 받는다면 틀린 추정

점추정은 오차를 필연적으로 동반한다는 약점이 있다.
따라서 점추정의 오차를 최소로 만들어야 바람직한 추정이라 할 수 있다.

점추정과 바람직한 점추정량의 조건

■ 바람직한 점추정량 조건

❶ 일치성 : 표본의 크기가 모집단 규모에 근접해야 한다.

일치성(consistency)은 표본이 모집단의 규모에 근접할수록 오차가 작아진다는 의미이다. 표본의 개수가 $n \to \infty$로 되어 모집단과 일치하면 오차는 0이 되므로, 표본이 커질수록 위험이 감소한다.

참고 일치성의 확인

$\lim_{n\to\infty}[E(\hat{\theta}_n - \theta)^2] = \lim_{n\to\infty} MSE(\hat{\theta}_n) = 0$ 이므로 추정량과 모수의 차이가 0이 되어 신뢰성이 커지게 된다.

점추정과 바람직한 점추정량의 조건

■ 바람직한 점추정량 조건

❷ 불편성 : 추정량이 모수와 같아야 한다.

추정량 $\hat{\theta}$로 모수 θ를 추정하여 $E(\hat{\theta}) = \theta$가 되면 가장 바람직한 추정이다.
이때의 추정량을 **불편추정량(unbiased estimator)**이라 한다.

$E(\hat{\theta}) \neq \theta$가 되면 추정량과 모수에 차이가있다는 의미이며, 이를 편의(biased)가 있다고 한다. 추정량에 대한 기대값이 모수와 동일하게 나타나면, 이는 표본추출이 오류가 나타날만한 영향이 없다는 불편성(unbiasedness)을 만족한다는 의미이다.

점추정과 바람직한 점추정량의 조건

■ 바람직한 점추정량 조건

❸ 유효성 : 추정량의 분산이 최소값이어야 한다.

유효성(efficiency)은 모수에 대한 추정량의 분산(분포)이 작을수록 추정량이 바람직하다는 미이다. 이러한 조건은 추정량이 여러 개일 경우, 이들을 서로 비교하여 가장 유효한 추정량을 확인할 때 필요하다.

만약 모수 θ에 대한 불편추정량이 $\hat{\theta}_1$과 $\hat{\theta}_2$로 두 개가 있다고 할 때,

$Var(\hat{\theta}_1) > Var(\hat{\theta}_2)$라면 $\hat{\theta}_1$ 보다 $\hat{\theta}_2$ 가 더 바람직한 추정량이라고 볼 수 있다.

점추정과 바람직한 점추정량의 조건

■ 바람직한 점추정량 조건

❹ 평균 오차제곱 : 평균 오차제곱이 최소값이어야 한다.

평균에서 측정치를 뺀 나머지를 오차라고 하는데, 이 차이가 최소여야 한다.

편차의 합은 0이 되므로 오차는 편차에 제곱을 한 **평균 오차제곱**(Mean Squared Error : MSE) $E[(\hat{\theta}-\theta)^2]$을 통해 살펴보며, 이 값이 최소가 되는 추정량이어야 한다.

즉 추정량과 모수의 차이가 최소가 되어야 한다.

점추정과 바람직한 점추정량의 조건

■ 바람직한 점추정량 조건

❺ 충분성 : 표본이 모집단의 대표성을 가져야 한다.

표본 $x_1, x_2, x_3, \cdots, x_n$으로부터 추정량 $\hat{\theta}$ 를 추정할 때,

확률함수를 $T(x_1, x_2, x_3, \cdots, x_n) = x_1$ 이라 하면 $\hat{\theta} = x_1$ 이 된다.

즉 확률함수에 어떤 값을 대입해도 x_1 만 도출되기 때문에

추정량 $\hat{\theta}$가 표본 $x_1, x_2, x_3, \cdots, x_n$의 정보를 모두 포함한다고 보기 어렵다.

표본은 모집단에 대해 대표성을 가져야 통계적인 의미가 있으므로,

$\hat{\theta} = T(x_1, x_2, x_3, \cdots, x_n)$ 의 정보가 모수 θ에 대한 모든 정보를 포함할 때, 추정량 $\hat{\theta}$ 를

모수 θ에 대한 충분한 추정이라 하고, 이를 충분성(sufficiency)이 확보되었다고 한다.

❶ 일치성, ❷ 불편성, ❸ 유효성, ❹ 평균오차제곱은 바람직한 추정량에 한정되는 조건

이지만 ❺ 충분성은 통계학 전체에 적용될 수 있는 일반적인 통계량에 관한 조건이다.

구간추정

■ 구간추정을 사용하는 이유

조사자의 입장에서 오차를 줄이기 위하여 명확한 수치를 제시하는 점추정 대신 신뢰도를 제시하면서 상한값과 하한값으로 모수를 추정하는 구간추정을 사용

■ 신뢰구간

상한값과 하한값의 구간으로 표시되며, 신뢰수준을 기준으로 추정된 점으로부터 음(-)의 방향과 양(+)의 방향으로 하한과 상한을 표시

구간추정

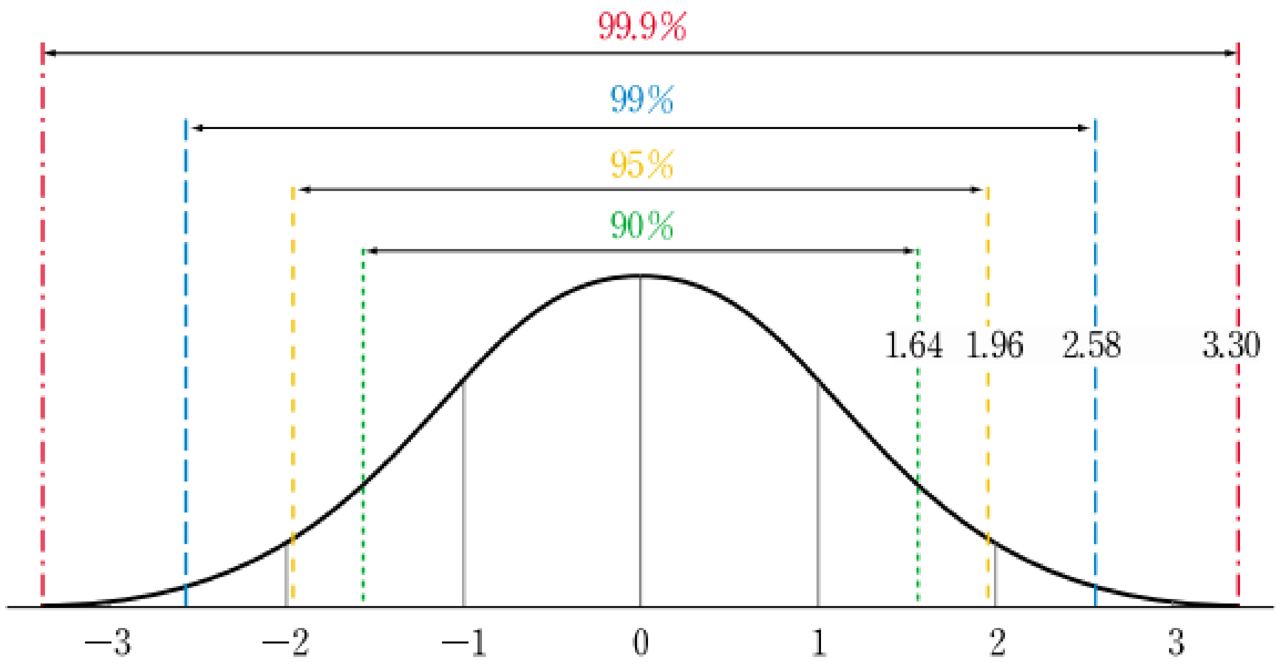
모평균(μ)을 추정할 때 표본평균을 $\bar{x}$

표준오차를 SE

표본평균 $\bar{x}$

모집단 평균에 대한 신뢰구간

$$\bar{x} - z \cdot SE \le \mu \le \bar{x} + z \cdot SE$$



예제 6-1 **구간추정**

토익 점수의 평균과 표준오차가 각각 (평균)=500, (표준오차)=100일 때, 신뢰도 90%, 95%, 99%에서 구간추정으로 평균에 대한 모수를 추정하라.

구간추정 (예제 풀이)

• 신뢰도 90%에서 구간추정

$\overline{x} = 500$, $SE = 100$, $z = 1.64$이므로 다음과 같이 구할 수 있다.

$$500 - 1.64 \cdot 100 \leq \mu \leq 500 + 1.64 \cdot 100$$
$$336 \leq \mu \leq 664$$

• 신뢰도 95%에서 구간추정

$\overline{x} = 500$, $SE = 100$, $z = 1.96$이므로 다음과 같이 구할 수 있다.

$$500 - 1.96 \cdot 100 \leq \mu \leq 500 + 1.96 \cdot 100$$
$$304 \leq \mu \leq 696$$

• 신뢰도 99%에서 구간추정

$\overline{x} = 500$, $SE = 100$, $z = 2.58$이므로 다음과 같이 구할 수 있다.

$$500 - 2.58 \cdot 100 \leq \mu \leq 500 + 2.58 \cdot 100$$
$$242 \leq \mu \leq 758$$

구간추정 (예제 풀이)

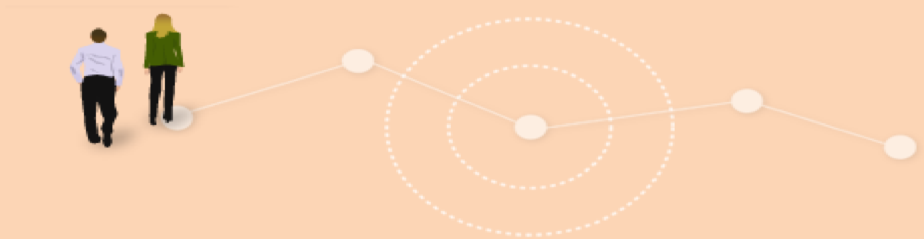
참고 신뢰도 100%

신뢰도를 100%로 하면 조사 결과가 모두 맞는 것이므로 최상의 결과가 될 것이다. 그러나 신뢰도 90%, 95%, 99%의 모평균 구간추정 결과에서 보는 바와 같이 신뢰도가 올라갈수록 구간은 점점 넓어진다. z값이 올라가기 때문에 그런 결과가 발생하는데, 100%에 해당하는 z값은 ∞다.

이는 $-\infty \leq \mu \leq \infty$의 값을 갖는다는 것이므로, 조사 결과가 틀릴 확률은 당연히 0%이지만 전혀 의미 없는 결과가 된다. 따라서 신뢰도를 약간 낮추더라도 유의미한 구간을 확인하는 게 더 유용할 수 있다. 즉, 조사자가 조사 대상에 따라 신뢰수준을 정하고 구간을 추정해야 한다.

02 모평균의 구간추정

:: Keywords 모수를 아는 경우의 z분포의 활용 | 모수를 모르는 경우의t분포 활용
표본의 크기 결정 | 모집단의 분산추정 | 모집단의 비율 추정
표본의 개수 측정



모평균의 구간추정

농구 선수들의 득점을 예측할 수 있을까?

[그림 6-2] 농구 선수의 득점 순간

농구는 경기 진행 속도도 빠르고 선수들의 행동도 아주 민첩하다. 또한 경기당 득점수도 개인별로 상당히 많은 차이가 난다.

그렇다면 어떤 한 선수를 특정하여 한 경기에서 득점하게 될 점수를 예측할 수 있을까? 물론 단 한 경기만 진행된다면 득점을 예측하기는 쉽지 않다. 하지만 프로 경기의 경우 시즌 내내 경기가 이어지므로, 과거의 경기 내용을 득점 추정의 근거로 활용할 수 있다.

이때 특정 선수가 지금까지 출전한 경기 내용만으로 단순한 그래프를 그려 추이를 예측한다면 과학적인 방법이라 할 수 없다. 과학적인 방법으로 득점을 예상하려면 통계를 적용해야 한다. 특정 선수가 지금까지 얻은 점수를 분석했을 때 일정한 규칙을 띤 분포를 구성하진 않더라도 이를 이용하여 신뢰수준을 적용하여 추정할 수 있으며, 이는 훨씬 신뢰할 만한 수치라 할 수 있다.

모평균의 신뢰구간 (모집단의 분산을 아는 경우)

■ 모집단의 분산을 아는 경우

모집단의 정보를 아는 경우는 거의 없음

→ 분산을 알고 있다고 가정하는 이유는

분산을 모르는 경우를 학습할 때 도움이 되기 때문

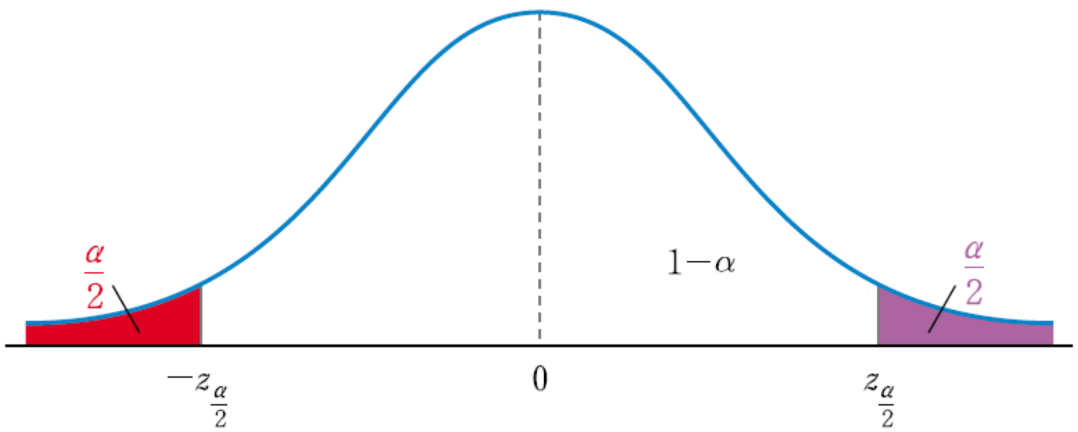
통계학에서는 90%, 95%, 99%, 99.9%에 해당하는 신뢰구간을 많이 이용

→ 신뢰구간에 속하지 않는 10%, 5%, 1%, 0.1%를 α라 했을 때,

신뢰구간은 1- α로 표현

모평균의 신뢰구간 (모집단의 표준편차를 아는 경우)

$$100(1-\alpha)\% = P\left(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}\right)$$



모집단의 표준편차를 알고 있으므로 평균이 μ, 표준오차가 $\frac{\sigma}{\sqrt{n}}$ 인 정규분포를 이룬다고 가정하면, 신뢰구간은

$$\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

모평균의 신뢰구간 (모집단의 표준편차를 아는 경우)

참고 **신뢰구간 $1-\alpha$의 도출 과정**

모집단에 대한 정보를 알고 있으며, 표본이 평균 μ, 표준오차 $\frac{\sigma}{\sqrt{n}}$ 인 정규분포를 이룬다고 가정하면, 신뢰구간을 다음과 같이 구할 수 있다.

$$-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}$$

$$\Rightarrow -z_{\frac{\alpha}{2}} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_{\frac{\alpha}{2}}$$

$$\Rightarrow -z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \bar{x}-\mu \le z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow -\bar{x}-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{x}+z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow \bar{x}-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

모평균의 신뢰구간 (모집단의 표준편차를 아는 경우)

예제 6-2 **모집단의 분산을 아는 경우 모평균의 신뢰구간 구하기** 준비파일 | 6장_신뢰구간.xlsx

A 회사에서 생산하는 전구의 평균 수명을 확인하고자 한다. 무작위로 200개의 표본을 추출하여 수명을 측정했더니 평균 수명이 30,000시간, 모분산은 250,000이었다. 이때 95%의 신뢰수준으로 모평균의 신뢰구간을 추정하라.

모평균의 신뢰구간 (예제 풀이)

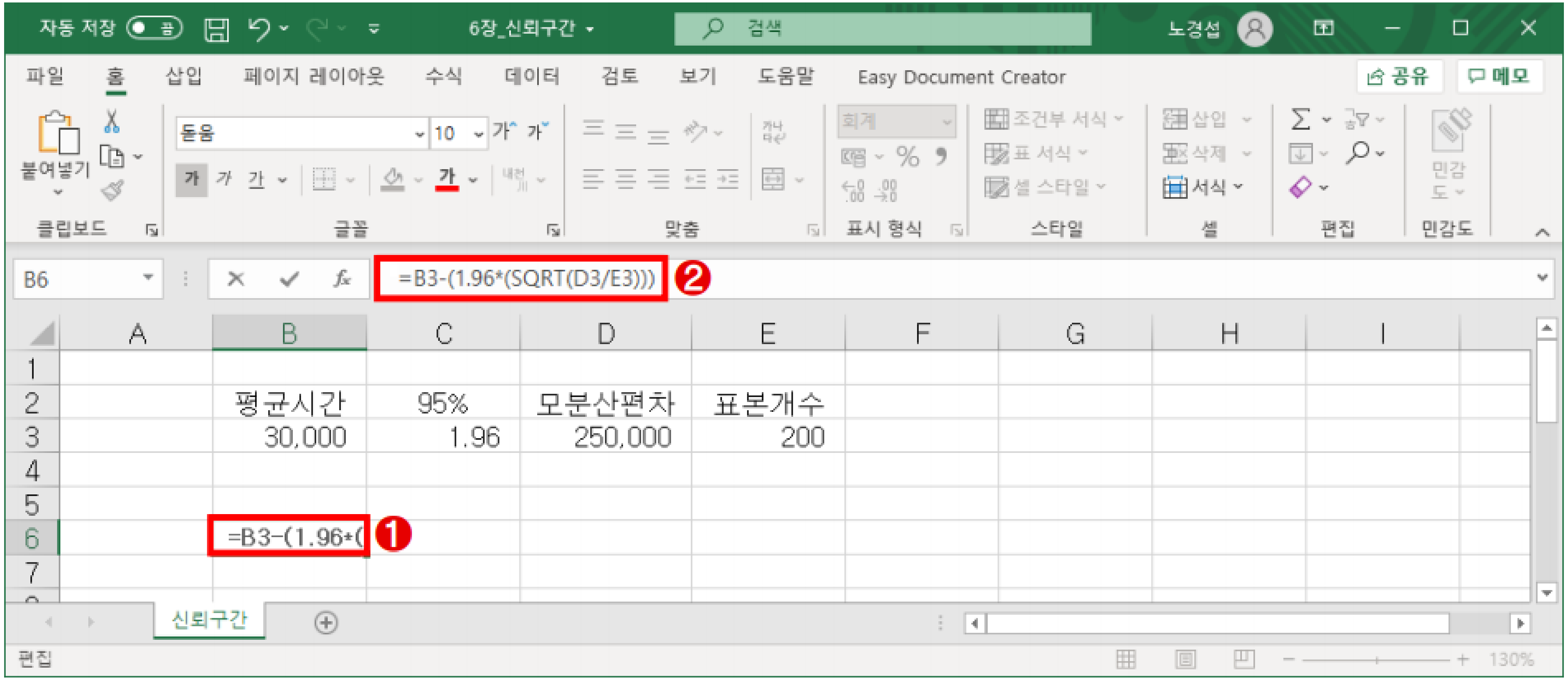
풀이

$\overline{x} = 30{,}000$, $z = 1.96$, $\sigma^2 = 250{,}000$, $n = 200$이므로 다음과 같이 구할 수 있다.

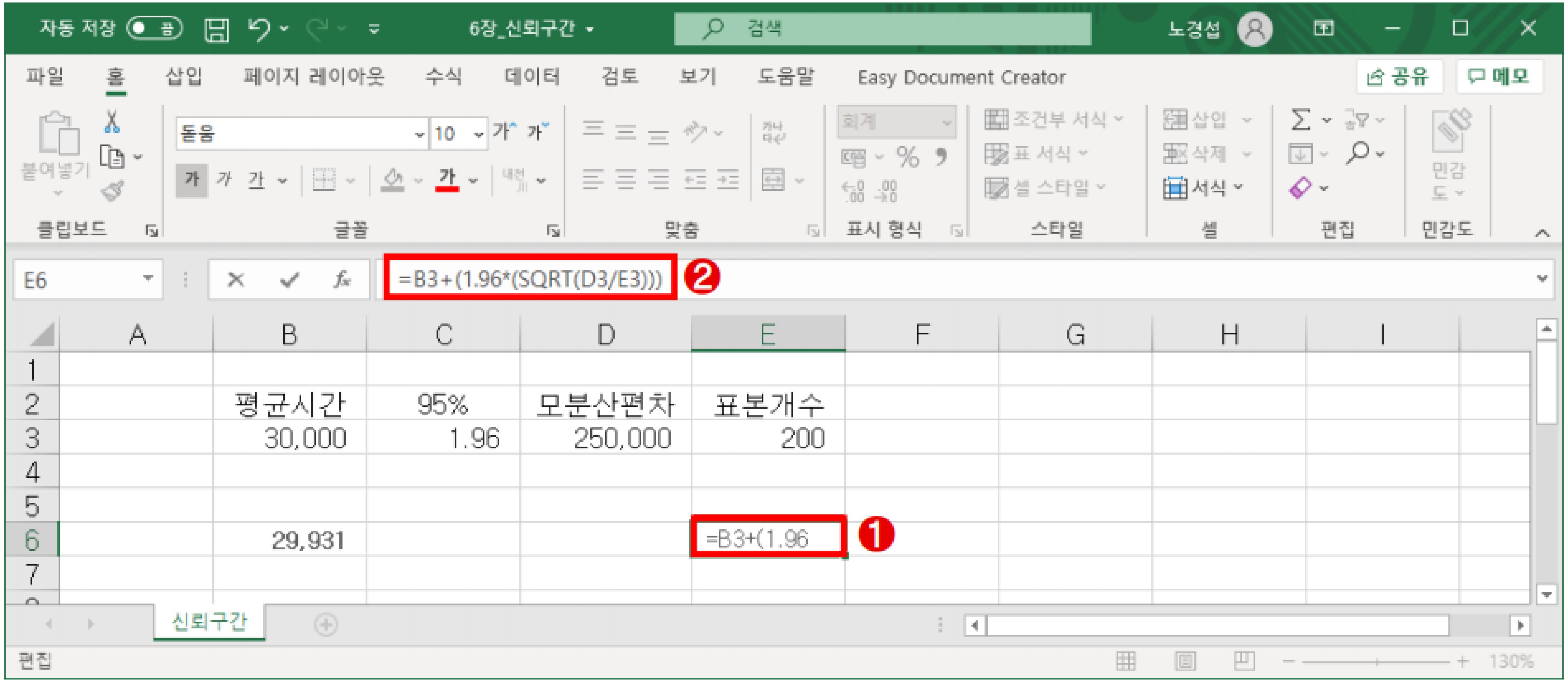
$$30{,}000 - 1.96 \cdot \sqrt{\frac{250{,}000}{200}} \le \mu \le 30{,}000 + 1.96 \cdot \sqrt{\frac{250{,}000}{200}}$$

$$29{,}931 \le \mu \le 30{,}069$$

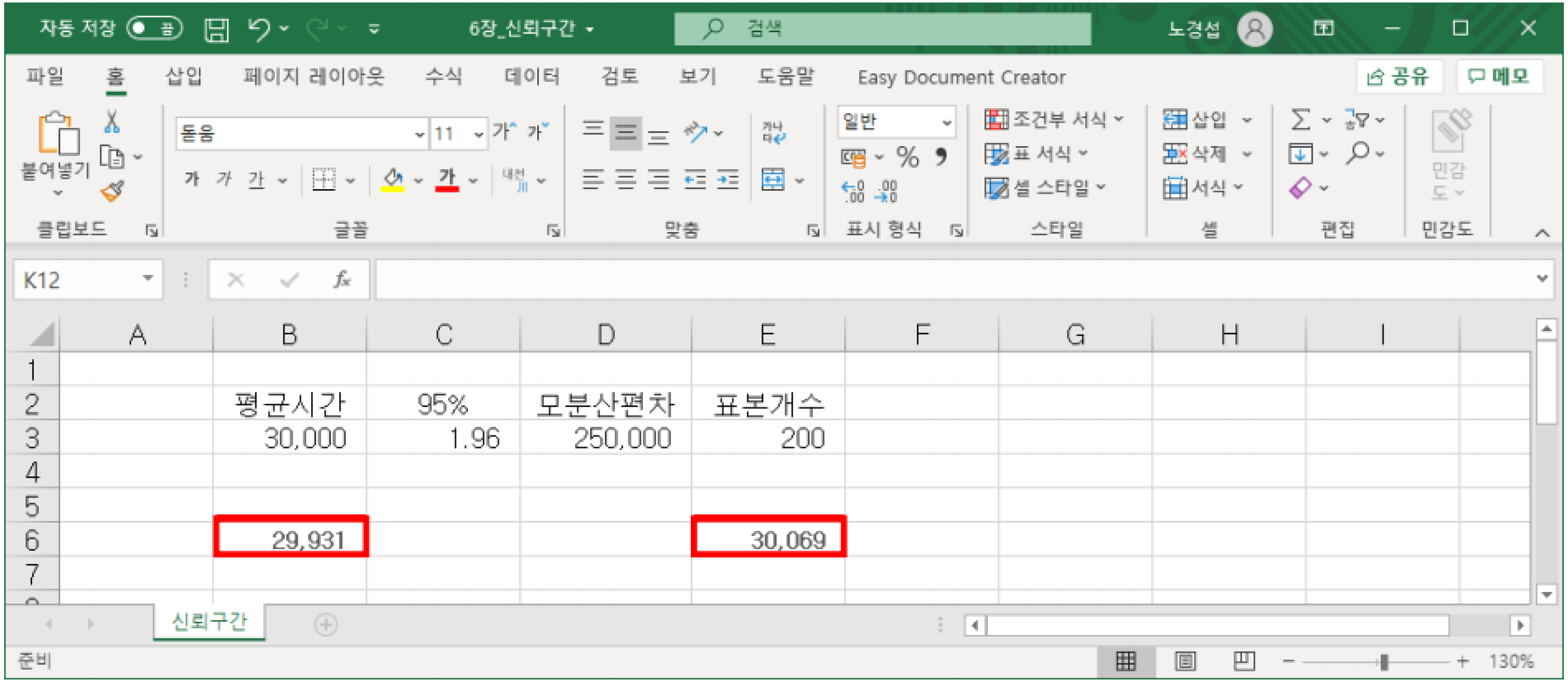
모평균의 신뢰구간 (예제 Excel 풀이)



모평균의 신뢰구간 (예제 Excel 풀이)



모평균의 신뢰구간 (예제 Excel 풀이 완성)



모평균의 신뢰구간

Note **표준편차 vs. 표준오차**

모집단에서 표본 100개를 추출한 후, 평균을 추정하는 실험을 한다고 가정해보자. 추출한 100개의 표본에 대해 평균($\bar{x}$)을 구한 후에 분포까지 확인해야 표본의 특성을 더 정확하게 파악할 수 있다. 분포를 표현하는 방법에는 **표준편차**(standard deviation)와 **표준오차**(standard error)가 있다.[5]

■ 표준편차

3장에서 중심경향도를 파악하고 분포를 나타내는 산포도를 확인하면서 분산과 표준편차를 살펴보았다. 표준편차는 표본평균으로부터 표본들이 흩어져 있는 산포를 나타내기 위해, 분산을 먼저 구한 후 다시 제곱근을 취한 값이다. 모평균을 알 수 없으므로 표본평균으로 모평균을 추정했으며, 표본의 분포를 확인하고자 표준편차를 구했다. 즉, 모평균을 추정하기 위해 표본을 추출해서 표본의 평균과 표본의 특성을 나타내는 것이 표준편차다. 표준편차는 다음과 같이 구한다.

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

모평균의 신뢰구간

■ **표준오차**

표준오차는 모평균을 추정하는 표본평균의 산포도를 나타낸다. 표본평균의 산포라는 말은 표본이 여러 개임을 의미한다. 단순히 1개의 평균으로 표본을 추출하기보다 2개 이상의 표본평균으로 모수를 추정한다면 더 정확하게 추정할 수 있다. 표본의 추출 횟수를 최대한 늘려서 $n \to \infty$로 할 수 있다면, 이들의 평균은 모수와 일치하게 될 것이다. 다시 말해, 모집단에서 k개의 표본을 추출해서 k개 $(\overline{x}_1, \overline{x}_2, \overline{x}_3, \cdots, \overline{x}_k)$의 평균을 구하면 1개의 표본으로 모평균을 추정할 때보다 모수에 더 근접하게 추정할 수 있다.

표준오차는 추출된 표본들의 숫자를 늘려서 평균을 구한 후, 이들 간의 표준편차를 나타낸 것이다. 표준오차는 평균들 간의 분포를 나타내므로 표준오차가 줄어들수록 평균을 나타내는 점들이 집중적으로 모여 있다. 따라서 이 경우 모수의 추정이 정확하게 이루어졌음을 판단할 수 있다. 평균의 표준오차는 다음과 같이 구한다.

$$\text{모분산을 아는 경우 : } S.E = \frac{\sigma}{\sqrt{n}}$$

$$\text{모분산을 모르는 경우 : } S.E = \frac{s}{\sqrt{n}}$$

모평균의 신뢰구간 (모집단의 표준편차를 모르는 경우)

■ 모집단의 표준편차를 모르는 경우

모집단의 정보를 아는 경우 → 공식에 대입하여 신뢰구간 추정

모집단의 표준편차를 모르는 경우

→표본의 표준편차를 이용해서 신뢰구간을 추정

표본의 분산 s^2과 표본의 표준편차 s는

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

모평균의 신뢰구간 (모집단의 표준편차를 모르는 경우)

표본의 표준편차를 이용한 신뢰구간은

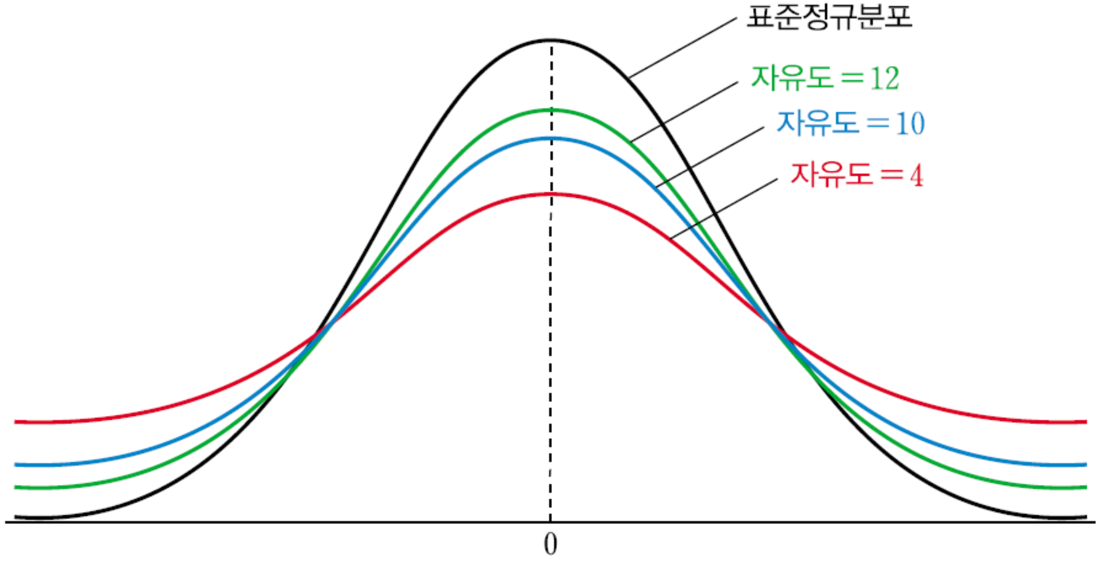
모표준편차를 이용한 신뢰구간보다 틀릴 가능성이 더 크므로,

신뢰구간의 범위가 더 커질 수밖에 없으므로 **t분포를 이용**

∵ t분포는 모수를 알지 못한 상황에서 정규분포를 이루는 모집단에서 추출한

표본의 크기가 작을 때의 추정과 검정에 사용

특히 t분포는 자유도에 따라 서로 다른 분포를 가짐.



모평균의 신뢰구간 (예제 풀이)

예제 6-3 **모집단의 분산을 모르는 경우 모평균의 신뢰구간 구하기** 준비파일 | 6장_신뢰구간2.xlsx

다음은 같은 반 학생 12명의 신장을 조사한 자료다. 학생들 신장에 대해 95% 신뢰수준으로 모평균의 신뢰구간을 추정하라.

번호	1	2	3	4	5	6	7	8	9	10	11	12
키	168	160	170	162	168	163	164	167	175	179	161	155

모평균의 신뢰구간 (예제 풀이)

풀이

학생들 신장의 평균과 표준편차를 구하면 다음과 같다.

$$\bar{x} = \frac{168 + 160 + \cdots + 161 + 155}{12} = 166$$

$$s = \sqrt{\frac{(168-166)^2 + (160-166)^2 + (170-166)^2 + \cdots + (155-166)^2}{11}} = 6.6469$$

이때 표준오차는 $\frac{s}{\sqrt{n}} = \frac{6.6469}{\sqrt{12}} = 1.9188$이고, 자유도는 $12-1=11$이다.

양측검정을 해야 하므로 자유도가 11이면서 95%의 신뢰구간에 해당하는 t값을 $\alpha = 0.025$에서 확인한다. t분포표에서 찾으면 95%의 $t_{\frac{\alpha}{2}} = 2.201$임을 알 수 있다. 따라서 신뢰구간을 구하면 다음과 같다.

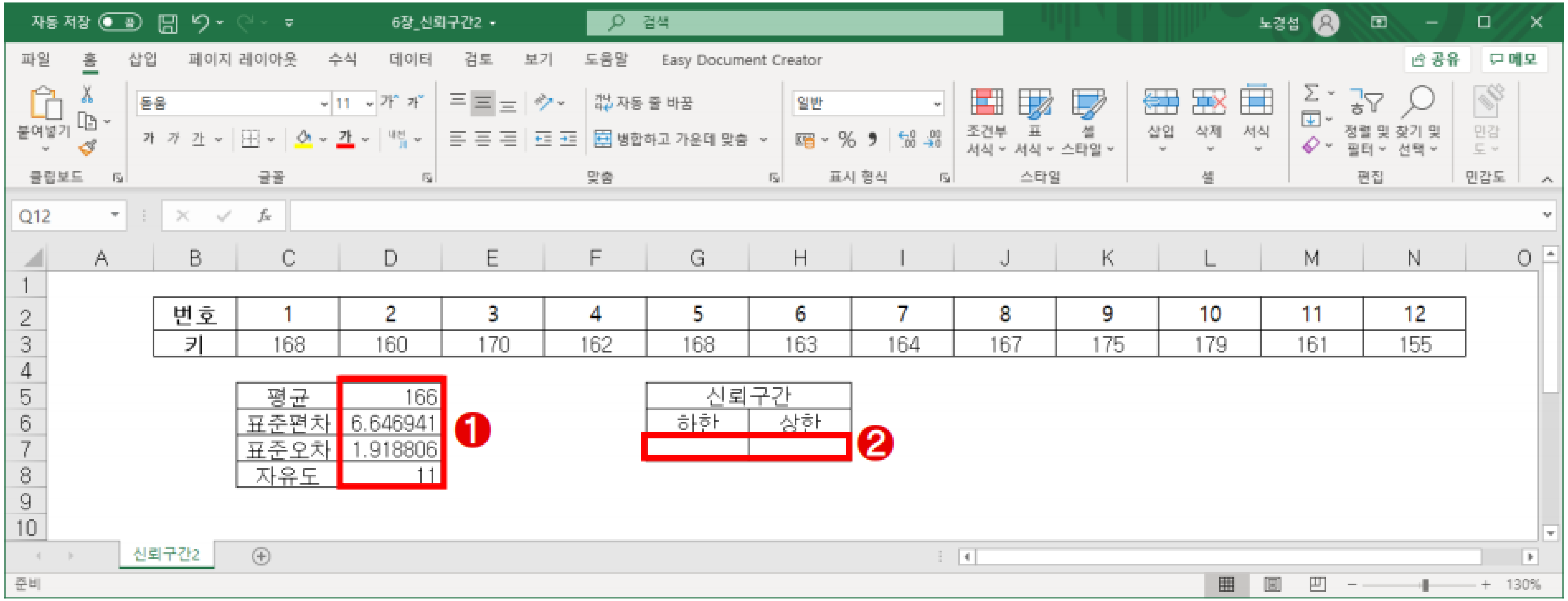
$$166 - 2.201 \cdot \frac{6.6469}{\sqrt{12}} \le \mu \le 166 + 2.201 \cdot \frac{6.6469}{\sqrt{12}}$$

$$166 - 2.2021 \cdot 1.9188 \le \mu \le 166 + 2.2021 \cdot 1.9188$$

$$161.7767 \le \mu \le 170.2233$$

즉, 95% 신뢰구간은 최저 161.7767cm에서 최대 170.2233cm로 계산되었다.

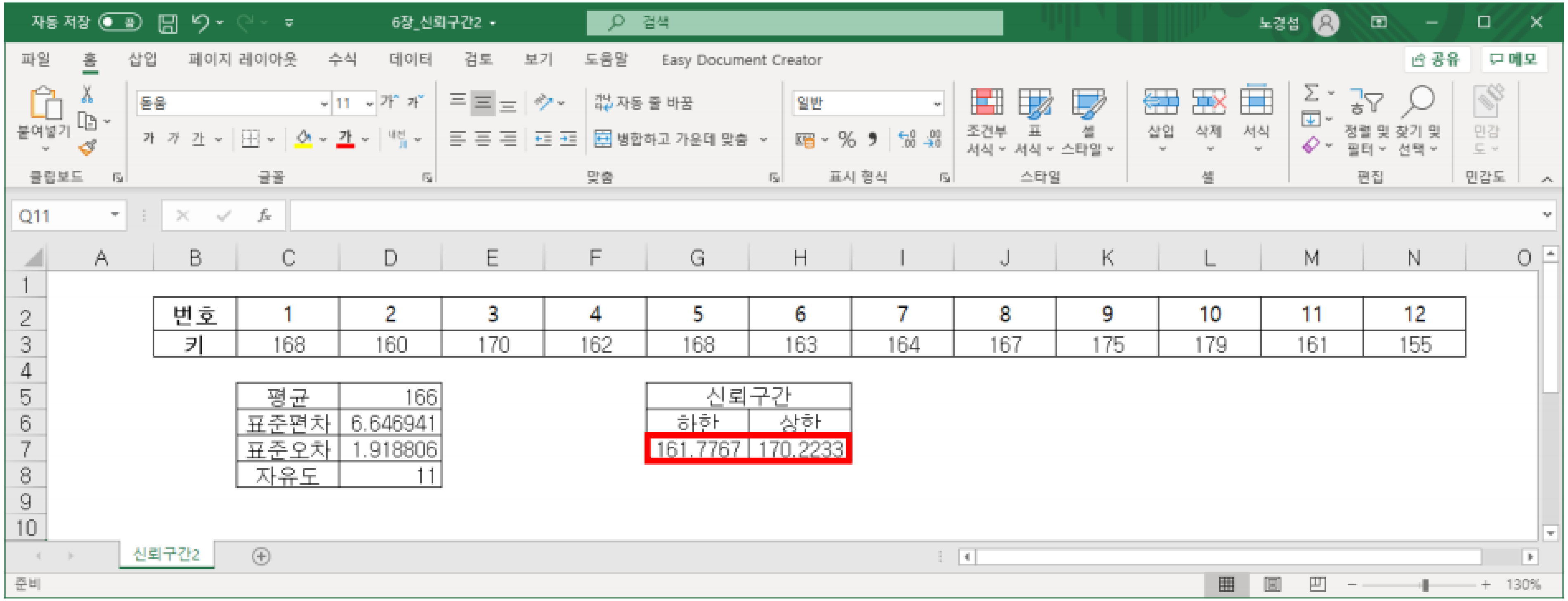
모평균의 신뢰구간 (예제 Excel 풀이)



❶ D5셀에 평균을 구하는 함수식 '=AVERAGE(C3:N3)'
D6셀에 표준편차를 구하는 함수식 '=STDEV.S(C3:N3)'
D7셀에 표준오차를 구하는 계산식 '=D6/SQRT(12)'
D8셀에 자유도 11을 입력

❷ G7셀에 하한값의 계산식 '=166−(2.2021*1.9188)'
H7셀에 상한값의 계산식 '=166+(2.2021*1.9188)' 을 입력한다.

모평균의 신뢰구간 (예제 Excel 풀이)



번호	1	2	3	4	5	6	7	8	9	10	11	12
키	168	160	170	162	168	163	164	167	175	179	161	155

평균	166
표준편차	6.646941
표준오차	1.918806
자유도	11

신뢰구간	
하한	상한
161.7767	170.2233

표본의 크기 결정

표본의 크기가 커진다는 것 → 통계량에 대한 신뢰도가 높아지는 것

표본의 크기가 작아진다는 것 → 모집단을 대표할만한 대표성과 신뢰도 ↓

그렇다면, 적절한 표본의 크기는?

$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot SE \quad \Rightarrow \quad \bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{d} \right)^2 \quad (d \ : \ \text{허용오차})$$

표본의 크기 결정

참고 허용오차의 도출 과정

$z_{\frac{\alpha}{2}} \cdot \dfrac{\sigma}{\sqrt{n}}$가 유의수준 내에서 인정되는 신뢰구간을 형성하므로 이를 허용오차 d라 할 수 있다. $\bar{x}$는 허용오차 100(1−α)% 확률의 신뢰구간에 포함된다. 따라서 $P(|\bar{x}-\mu| \leq d) = 1-\alpha$($d$: 허용오차)로 표현할 수 있다. 여기서 n을 도출하면 다음과 같다.

$$d = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \;\Rightarrow\; d \cdot \sqrt{n} = z_{\frac{\alpha}{2}} \cdot \sigma \;\Rightarrow\; \sqrt{n} = \frac{z_{\frac{\alpha}{2}} \cdot \sigma}{d} \;\Rightarrow\; \therefore\; n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{d} \right)^2$$

표본의 크기 결정

예제 6-4 **적절한 표본 개수 구하기** 준비파일 | 6장_표본개수.xlsx

주유소의 주유기가 정확한 용량으로 주유되는지를 확인하기 위해 1리터씩 용량을 재는 조사를 하고자 한다.

(a) 최소 몇 개의 표본을 검사해야 하는지 구하라.
(단, 신뢰수준 99%, 허용오차 ±100ml, 표준편차 150ml)

(b) 알려진 표준편차가 없어서 10개의 표본을 추출해 (표준편차) = 170ml를 계산했다.
신뢰수준 95%, 허용오차 ±100ml일 때, 몇 개의 표본으로 조사해야 하는지 구하라.

표본의 크기 결정 (예제 풀이)

풀이

(a) $1-\alpha=0.99$이므로 $z_{\frac{\alpha}{2}}=2.58$, $d=100$, $s=150$이다. 따라서 표본 개수를 구하면 다음과 같다.

$$n=\left(\frac{z_{\frac{\alpha}{2}}\cdot\sigma}{d}\right)^2=\left(\frac{2.58\cdot 150}{100}\right)^2=14.9769$$

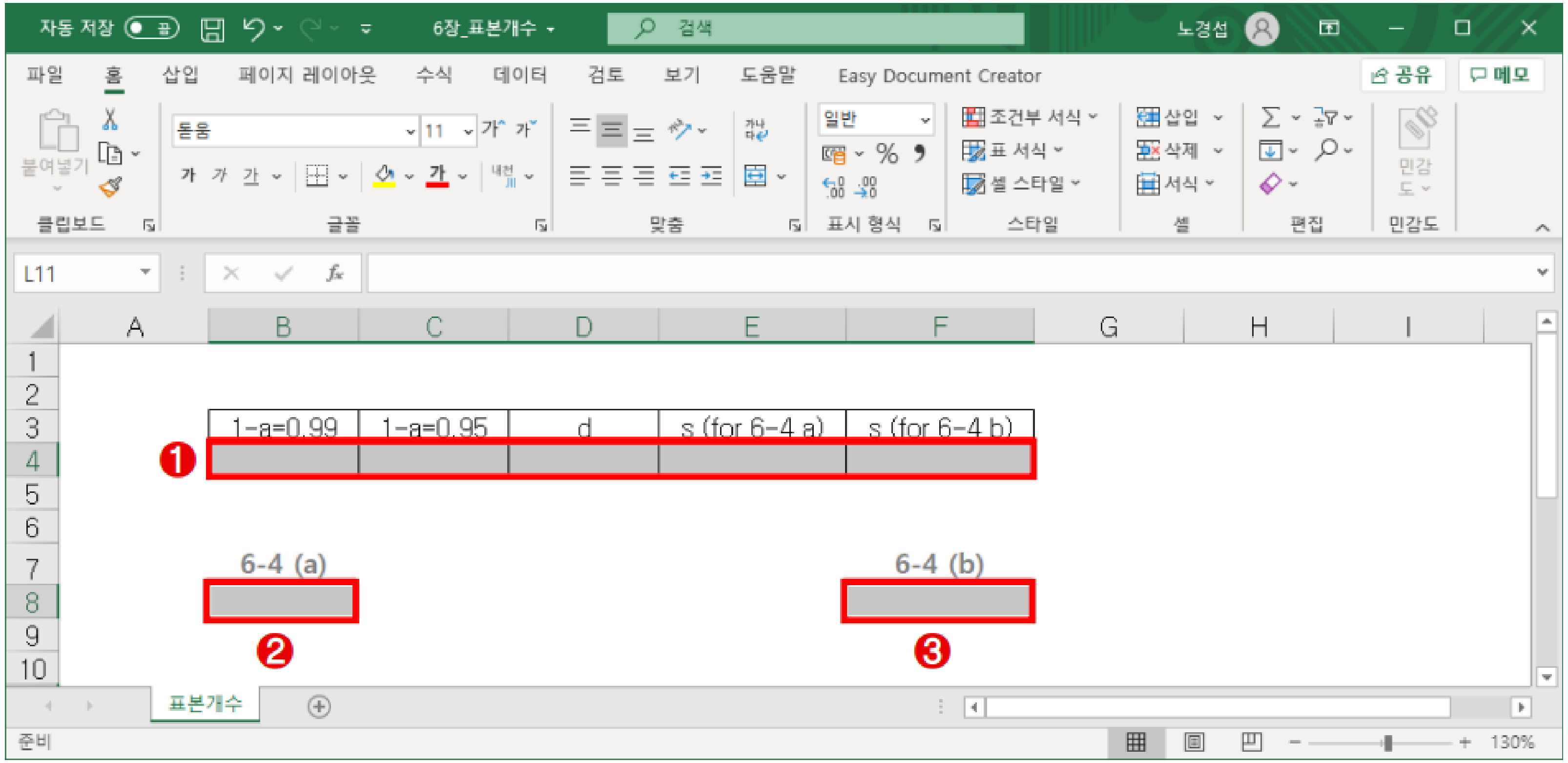
그러므로 최소한 15개의 표본으로 용량을 확인해야 한다.

(b) $1-\alpha=0.95$이므로 $z_{\frac{\alpha}{2}}=1.96$, $d=100$, $s=170$이다. 따라서 표본 개수를 구하면 다음과 같다.

$$n=\left(\frac{z_{\frac{\alpha}{2}}\cdot\sigma}{d}\right)^2=\left(\frac{1.96\cdot 170}{100}\right)^2=11.1022$$

그러므로 최소한 12개의 표본으로 용량을 확인해야 한다.

표본의 크기 결정 (예제 Excel 풀이)

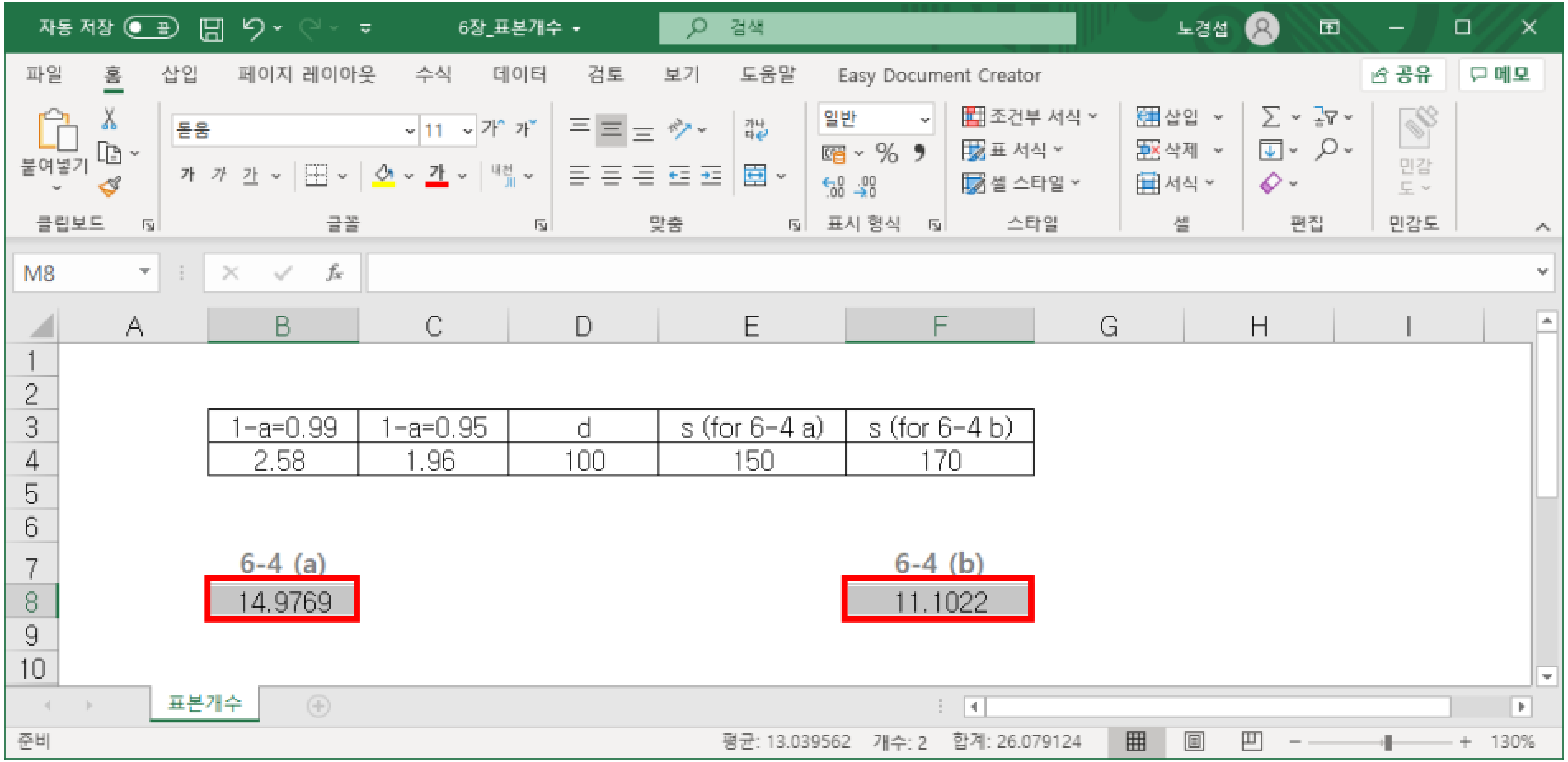


❶ B4셀에 ‘2.58’, C4셀에 ‘1.96’, D4셀에 ‘100’,
E4셀에 (a)의 표준편차인 ‘150’, F4셀에 (b)의 표준편차인 ‘170’을 각각 입력

❷ B8셀에 (a)를 구하는 수식 ‘=((B4*E4)/D4)^2’을 입력

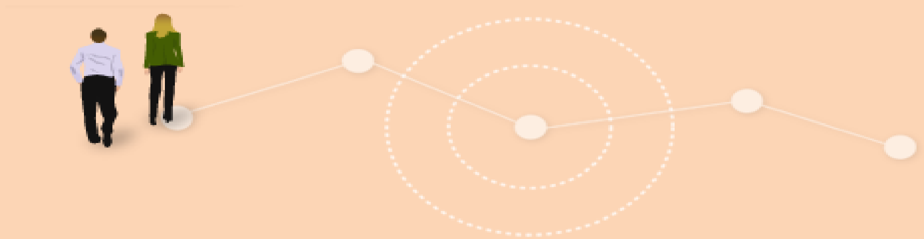
❸ E8셀에 (b)를 구하는 수식 ‘=((C4*F4)/D4)^2’을 입력

표본의 크기 결정 (예제 Excel 풀이 완성)



03 모집단 비율 및 분산의 구간추정

:: Keywords 모집단 비율의 신뢰구간 | 모집단 비율에서 표본의 크기 결정
모집단 분산의 신뢰구간



모집단 비율 및 분산의 구간추정

우리 학교 졸업생의 취업률은 얼마나 될까?

[그림 6-12] 구직

현재 대학생들의 가장 큰 관심사는 단연 취업이다. 보도에 따르면 청년 실업률이 대략 10%라는데, 이는 10명 중 9명이 취업을 했다는 말이다. 그런데 취업을 준비하는 학생들에게는 이러한 수치가 피부에 와 닿지 않는다. 실제로 느껴지는 취업률은 통계 결과로 나타난 수치보다 상당히 낮게 느껴진다.

졸업생의 취업률을 조사할 때 졸업생 명단을 받아서 전수조사를 하면 정확한 모수를 파악할 수 있을 것이다. 그러나 졸업생 전수조사는 만만한 일이 아니다. 따라서 표본을 대상으로 조사하여 취업에 성공한 비율을 추정하게 되는데, 이때 신뢰구간을 적용하여 구간추정을 하면 점추정보다 정확하게 추정할 수 있다.

모집단 비율의 구간추정

■ 모집단 비율의 구간

표본비율 $\hat{p}$에 대한 신뢰구간을 의미

→ 표본비율은

표본을 추출했을 때, 표본의 개수 n 중에서 특정 사건 t가 발생하는 빈도를 비율로 나타낸 것

$$\text{표본비율}(\hat{p}) = \frac{\text{특정 사건}(t)}{\text{표본의 개수}(n)}$$

표본이 충분히 클 때 표본비율의 신뢰구간은

$$\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}} \le p \le \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

$$\text{표준오차}(\sigma_{\hat{p}})\text{는} \quad \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

모집단 비율의 구간추정

참고 모집단 비율에서 표준오차 도출 과정

추출한 표본 중에서 특정한 사건이 '발생했다' 혹은 '발생하지 않았다'를 나타내는 분포는 이항분포다. 이항분포의 평균은 $\mu = E(x) = p$이고, 분산은 $V(x) = p(1-p)$이다.

계산된 분산은 표본평균과 모평균 사이에 어느 정도의 차이가 있는지를 판단할 수 있는 지표여야 한다. 따라서 표본의 개수로 나누고 제곱근을 취해 표본평균이 모평균으로부터 어느 정도 떨어져 있는지를 판단한다. 결국 표준오차($\sigma_{\hat{p}}$)는 $\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}$으로 표현할 수 있다.[8] 즉, 표준편차가 하나의 표본에서 각각의 개체에 대한 값들의 차이를 의미한다면, 표준오차는 표본을 통해 도출된 평균값이 모집단의 평균과 어느 정도의 차이가 발생하는지를 보여준다.

모집단 비율의 구간추정

예제 6-5 모집단 비율의 신뢰구간 구하기

준비파일 | 6장_모비율의 신뢰구간.xlsx

주유소의 주유 용량이 정확한지 확인하기 위해 국내 주유소 중 무작위로 77곳을 선정하여 용량을 계량했다. 그 결과 5곳의 주유소에서 주유해야 할 용량과 실제 주유된 용량에 차이가 있었다. 전체 주유소의 주유 용량에 대해 95% 신뢰수준으로 주유해야 할 용량과 실제 주유된 용량에 차이가 있을 비율에 대한 신뢰구간을 구하라.

모집단 비율의 구간추정 (예제 풀이)

77곳의 주유소 표본 중 5곳에서 용량의 차이

$\hat{p} = \frac{5}{77} = 0.0649$, $n = 77$, $\hat{p} = 0.0649$, $z_{\frac{\alpha}{2}} = 1.96$ 이므로

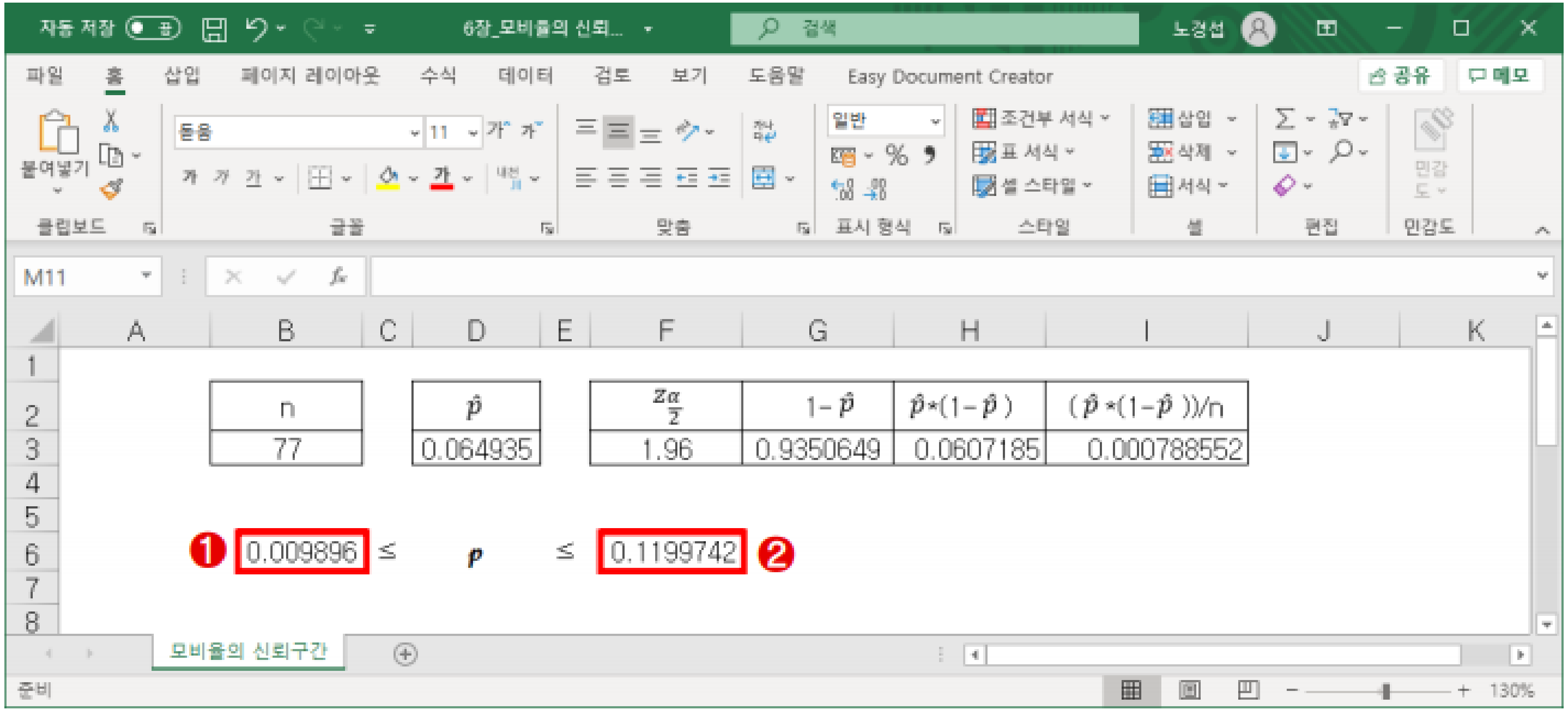
신뢰구간은

$$0.0649 - 1.96 \cdot \sqrt{\frac{0.0649 \cdot (1 - 0.0649)}{77}} \leq p \leq 0.0649 + 1.96 \cdot \sqrt{\frac{0.0649 \cdot (1 - 0.0649)}{77}}$$

$$0.009896 \leq p \leq 0.1199742$$

그러므로 전체 주유소의 0.98%~12.00%에서 주유해야 할 용량과
실제 주유된 용량에 차이가 발견될 수 있다.

모집단 비율의 구간추정 (예제 Excel 풀이)



❶ B6셀에 하한값의 계산식 '=D3−F3*SQRT(I3)'를 입력하고,

❷ F6셀에 상한값의 계산식 '=D3+F3*SQRT(I3)'를 입력

모집단 비율

■ 표본의 크기 결정

모집단의 비율을 추정하는 경우에도 표본의 크기를 어느 정도로 해야 할지 가늠해야 하는데, 신뢰구간의 오차한계를 어느 정도로 할 것인지를 먼저 정하면 적절한 표본의 개수를 결정할 수 있다.

→ 표본비율에서 상한과 하한을 결정하는 $z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$ 가

오차한계 d를 넘지 않아야…

▶표본의 개수 n은

$$d = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

$$d \cdot \sqrt{n} = z_{\frac{\alpha}{2}} \cdot \sqrt{\hat{p} \cdot (1-\hat{p})}$$

$$n = \hat{p} \cdot (1-\hat{p}) \cdot \left(\frac{z_{\frac{\alpha}{2}}}{d}\right)^2$$

모집단 비율

예제 6-6 모집단 비율에 대한 표본 크기 구하기 준비파일 | 6장_표본크기 결정.xlsx

휴대폰을 제조하는 회사에서 화면이 클수록 스마트폰 사용자의 만족도가 올라가는지를 조사하려고 한다. 큰 화면을 사용하면 만족도가 올라간다는 이용자의 비율에 대해 95%의 신뢰구간으로 조사하려고 할 때, 오차한계가 5% 이하인 표본의 크기를 구하라. (단, $\hat{p} = 0.95$)

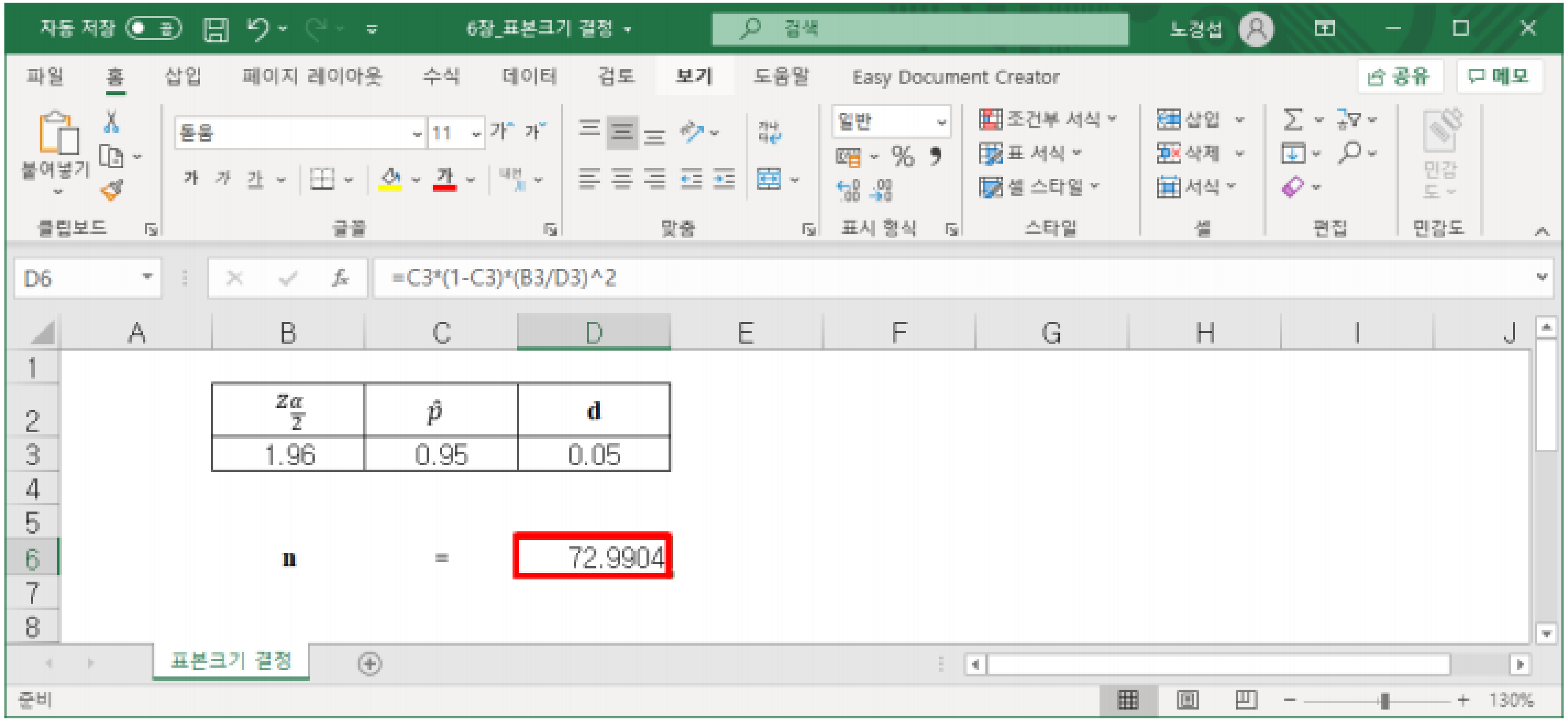
모집단 비율 (예제 풀이)

$z_{\underline{\alpha}} = 1.96, \quad \hat{p} = 0.95, \quad d = 0.05$ 이므로 표본의 개수는

$$n = 0.95 \cdot 0.05 \cdot \left(\frac{1.96}{0.05} \right)^2 = 72.99$$

따라서 최소 73명을 표본으로 설정해야 한다.

모집단 비율 (예제 Excel 풀이)



D6셀에 표본의 크기를 구하는 계산식 ‘=C3*(1−C3)*(B3/D3)^2’를 입력

모집단 분산의 구간추정

■ 모집단 분산의 구간

분산(혹은 표준편차)은 평균과 비율로는 알 수 없는 분포의 특성을 설명해주기에, 의사결정을 하는 데 중요한 기준이 될 수 있다.

분산을 분포로 나타낸 것이 χ^2분포

→ 정규분포를 이루는 모집단이라도 분산의 분포는 χ^2분포로 나타남.

분산이 σ^2 인 모집단으로부터 n의 표본을 추출해서 표본분산 s^2 를 계산하면,

자유도가 $(n-1)$인 χ^2 분포를 따르며

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$ 으로 나타낸다.

모집단 분산의 구간추정

X^2 분포는 정규분포의 모양을 따르지 않기 때문에

하한과 상한을 동일하게 표현하지 않고, 신뢰구간을 나타내면

$$\chi^2_{1-\frac{\alpha}{2}} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2}} \quad \Rightarrow \quad \frac{\chi^2_{1-\frac{\alpha}{2}}}{(n-1)s^2} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{\frac{\alpha}{2}}}{(n-1)s^2}$$

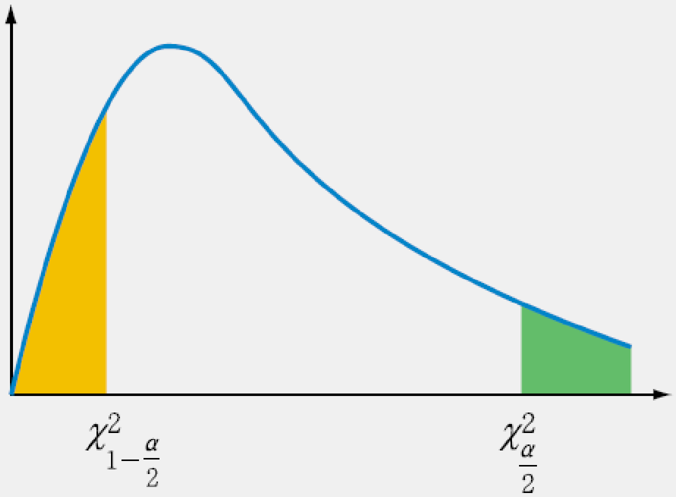
따라서 모집단 분산의 신뢰구간은

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

모집단 분산의 구간추정

Note χ^2**분포에서 상한과 하한을 다르게 표현하는 이유는?**

정규분포로 가정하면 신뢰구간의 상한과 하한을 $\overline{X} \pm Z_{\frac{\alpha}{2}} \cdot SE$ 와 같이 ±를 이용하여 중심으로부터 동일한 위치를 계산하면 된다. 하지만 χ^2분포는 분산 자체가 어느 정도 산포하고 있는지를 나타내는 분포다. 분산이 제곱값이므로 χ^2분포에서는 음수(−)가 나올 수 없다. 즉, 양수(+) 값만 존재한다. 따라서 좌우가 서로 대칭인 정규분포가 아니다.[9]



[그림 6-15] χ^2 분포

또한 α에 해당하는 영역은 오른쪽 부분의 영역을 나타낸다. 그래프의 전체 면적이 1이므로 좌측 영역은 $1-\alpha$가 된다. 신뢰구간에서 α는 모수를 포함하지 않을 확률이므로 상한과 하한을 $\chi^2_{\frac{\alpha}{2}}$과 $\chi^2_{1-\frac{\alpha}{2}}$으로 구분할 수 있다. 그러므로 모분산($\sigma^2$)을 추정하는 공식으로 다음 식을 사용한다.

$$\chi^2_{1-\frac{\alpha}{2}} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2}}$$

모집단 분산의 구간추정

예제 6-7 모집단 분산의 신뢰구간 구하기

준비파일 | 6장_모집단 분산의 신뢰구간.xlsx

학생들의 스마트폰 중독 정도를 확인하고자 스마트폰을 이용하는 초·중고생 50명을 대상으로 1일 평균 이용 시간을 조사했다. 그 결과 학생들은 스마트폰을 평균 4.3시간 이용하고 있으며, 분산이 2.5시간으로 나타났다. 스마트폰의 이용 시간에 대한 분산을 확인하기 위해 신뢰수준 95%에서 신뢰구간을 구하라.

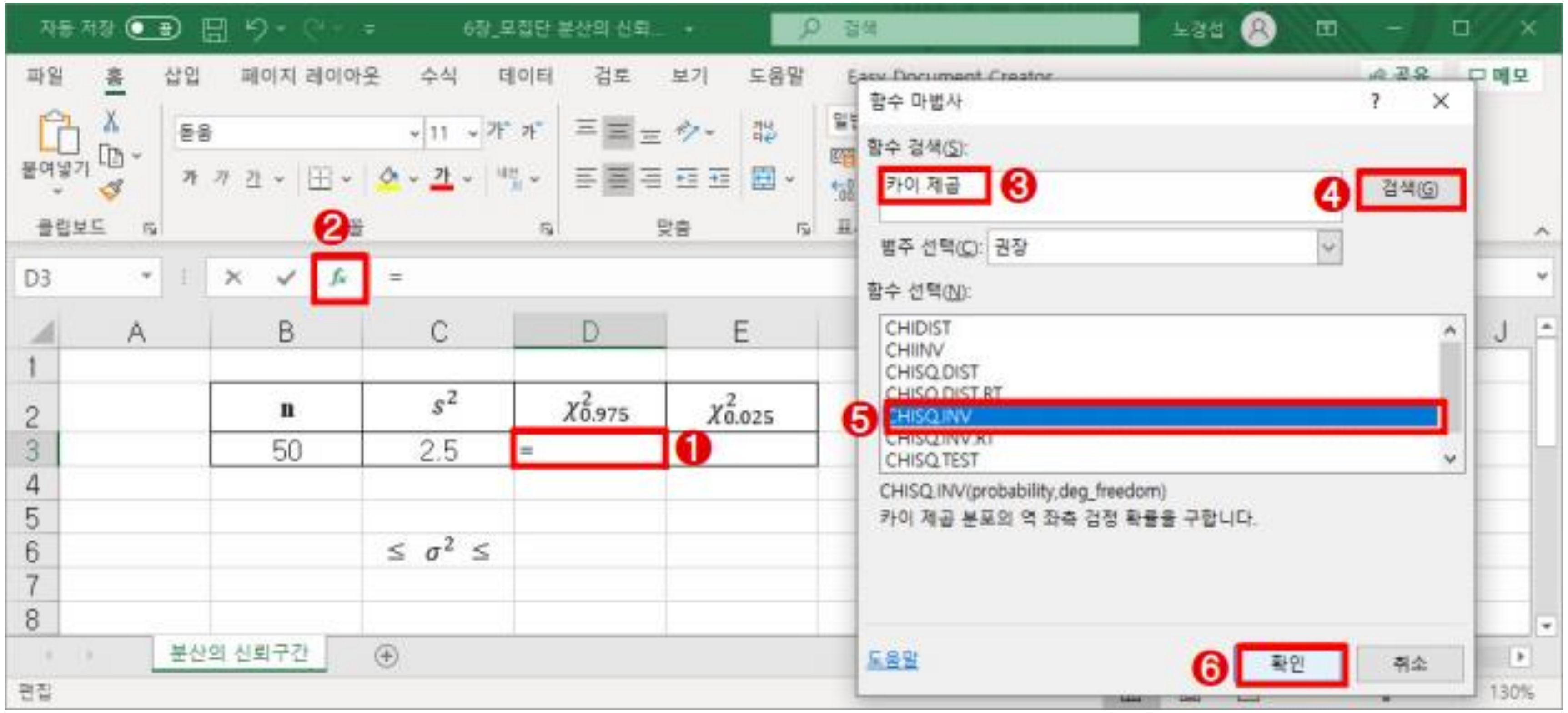
모집단 분산의 구간추정 (예제 풀이)

$n = 50$, $s^2 = 2.5$, $\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.975} = 70.22$, $\chi^2_{\frac{\alpha}{2}} = \chi^2_{0.025} = 31.55$ 이므로

신뢰구간은 $\dfrac{49 \cdot 2.5}{70.22} \le \sigma^2 \le \dfrac{49 \cdot 2.5}{31.55}$

$$1.74 \le \sigma^2 \le 3.88$$

모집단 분산의 구간추정 (예제 Excel 풀이)

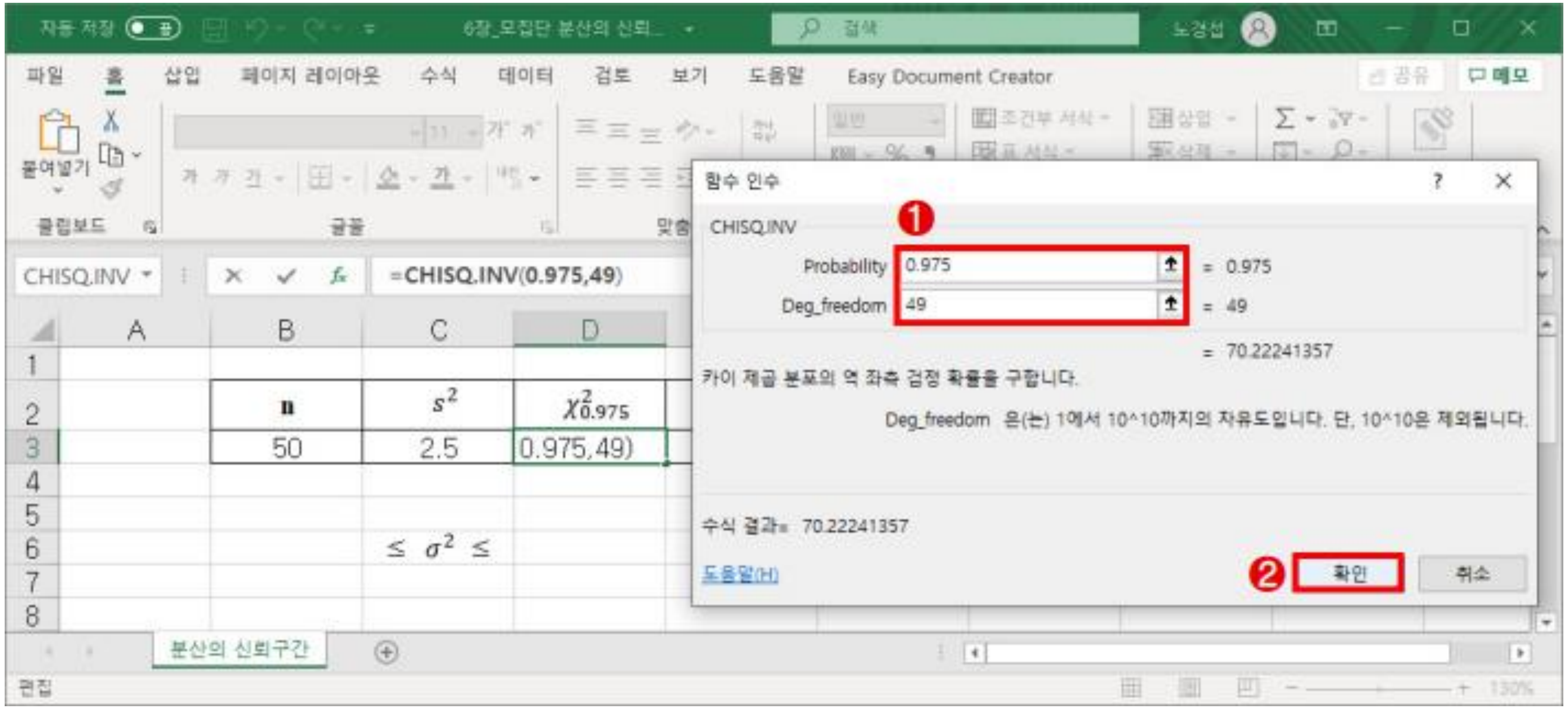


❶ D3셀로 셀 포인터를 이동한 후 ❷ fx 를 클릭한다. **[함수 마법사]** 창에서

❸ '함수 검색(S):'에 '카이제곱'을 입력하고 ❹ **[검색]**을 클릭

❺ '함수 선택(N):'에서 'CHISQ.INV' 함수 를 선택한 후 ❻ **[확인]**을 클릭

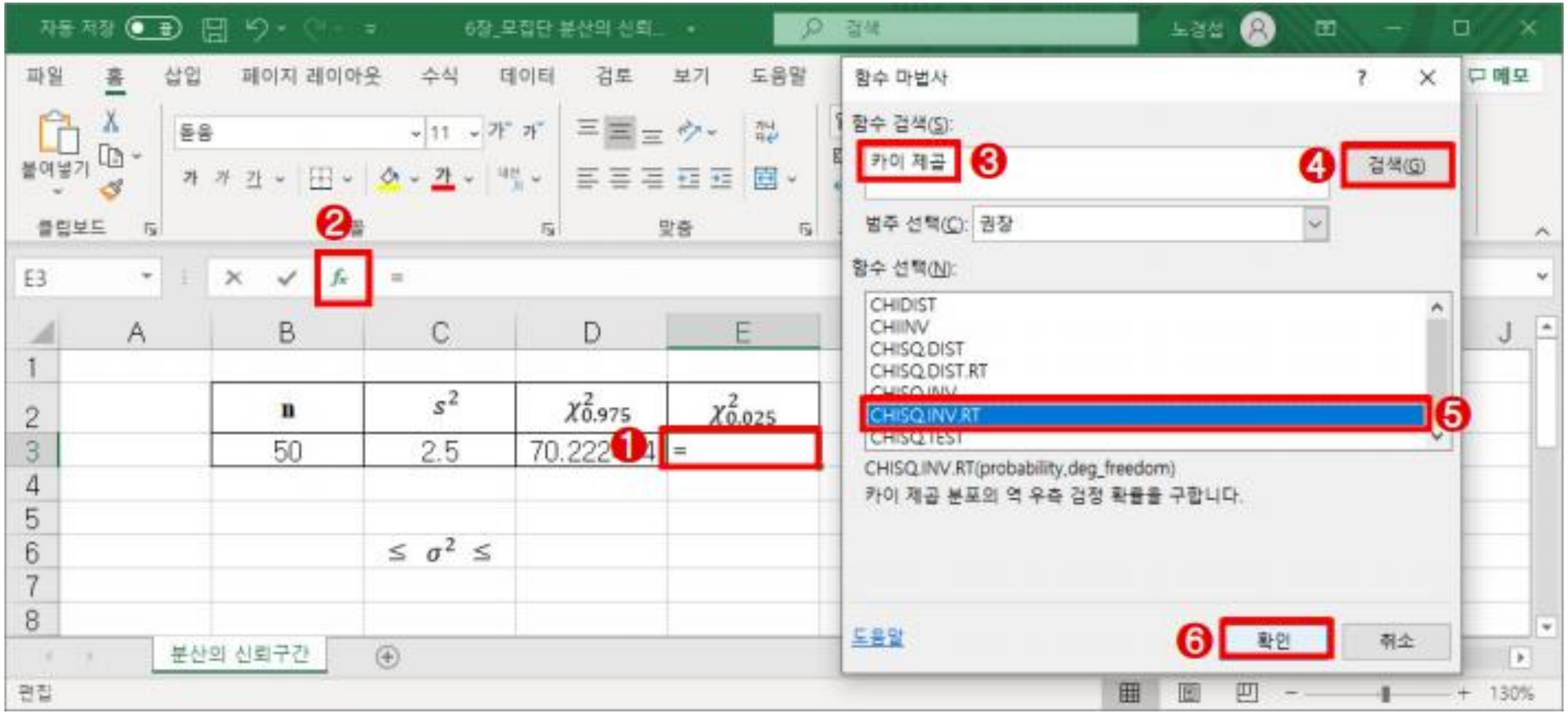
모집단 분산의 구간추정 (예제 Excel 풀이)



❶ 'Probability'에 0.975를 입력

❷ 자유도를 나타내는 'Deg_freedom'에 49를 입력한 후 ❸ [확인]을 클릭

모집단 분산의 구간추정 (예제 Excel 풀이)

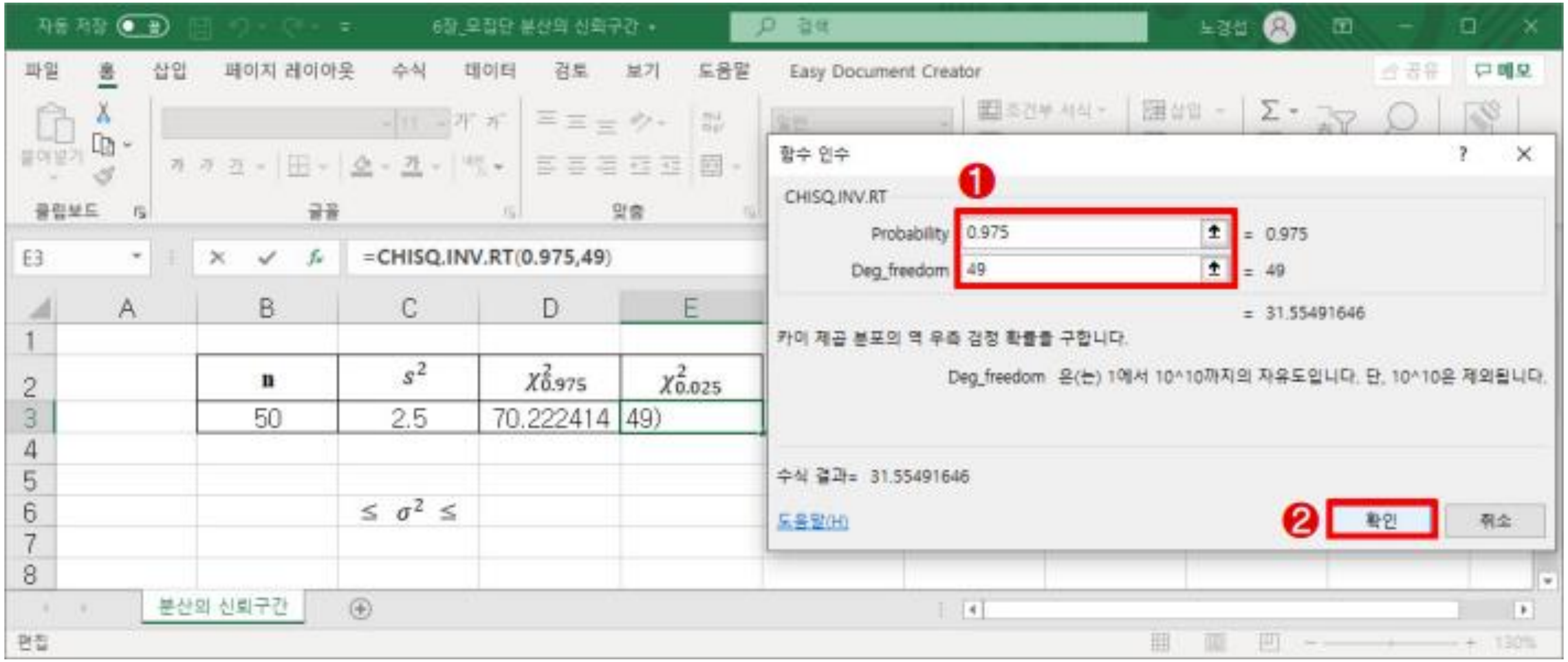


❶ E3셀로 셀 포인터를 이동 ❷ f_x 를 클릭

❸ '함수 검색(S):'에 '카이제곱'을 입력하고 ❹ [검색]을 클릭

❺ '함수 선택(N):'에서 'CHISQ.INV.RT' 함수를 클릭한 후 ❻ [확인]을 클릭

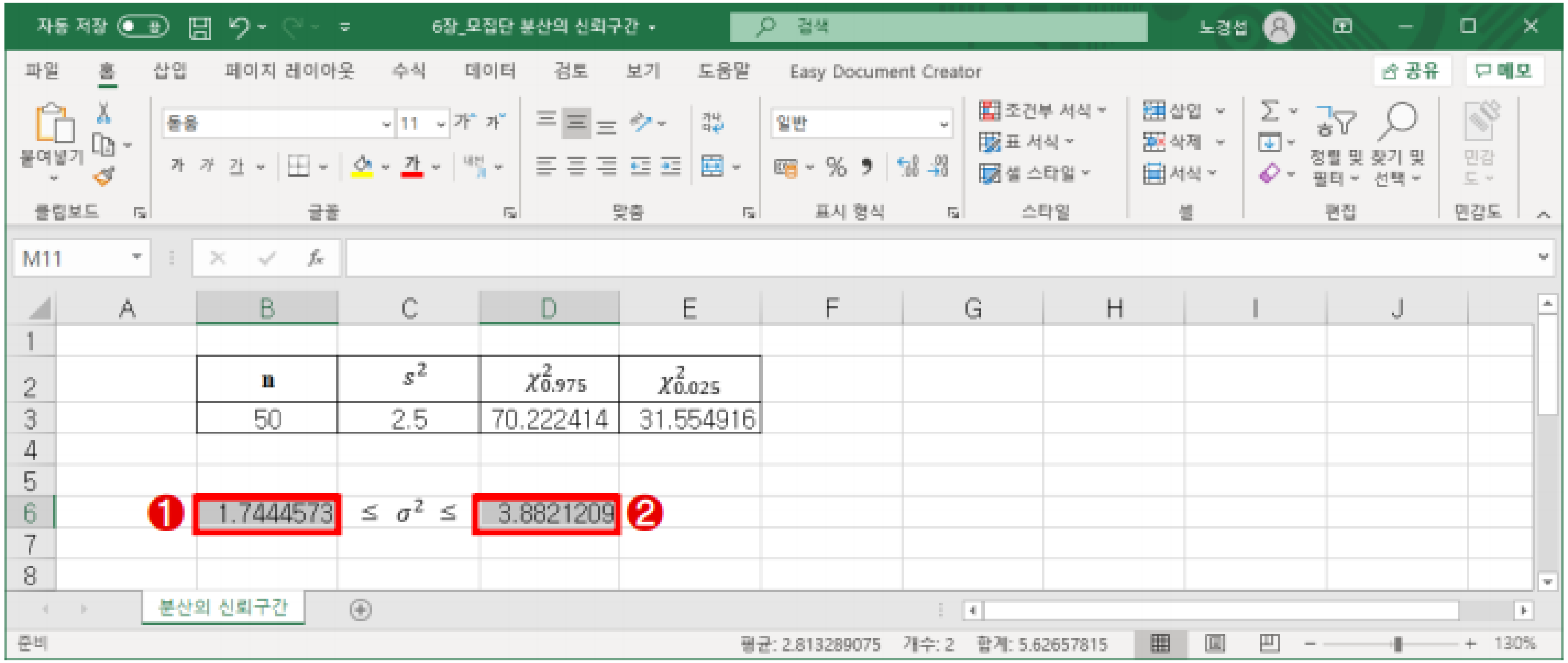
모집단 분산의 구간추정 (예제 Excel 풀이)



❶ 'Probability'에 0.975를 입력하고,

❷ 'Deg_freedom'에 49를 입력한 후 ❸ [확인]을 클릭

모집단 분산의 구간추정 (예제 Excel 풀이)



❶ B6셀에 하한값의 계산식 '=((B3−1)*C3)/D3'를 입력하고,

❷ D6셀에 상한값의 계산식 '=((B3−1)*C3)/E3'를 입력

Q&A

통계학, 제대로 시작하자!

통계의 쓰임을 이해하고,
실제로 활용할 수 있어야 한다.