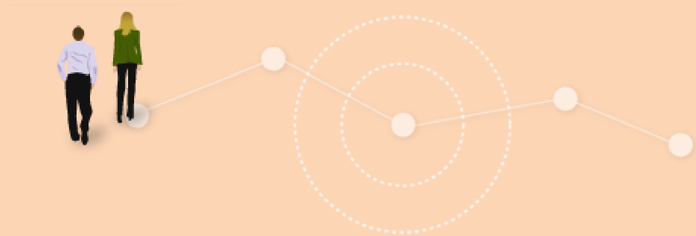
6 장

목 차

01 점추정과 구간추정

:: Keywords 점추정 | 구간추정 | 신뢰구간



추정이란

■ 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

■ 바람직한 점추정량 조건

➊ 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) $E[(\hat{\theta}\ \text{-}\theta)^2]$을 통해 살펴보며, 이 값이 최소가 되는 추정량이어야 한다.

즉 추정량과 모수의 차이가 최소가 되어야 한다.

점추정과 바람직한 점추정량의 조건

■ 바람직한 점추정량 조건

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 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

❹ 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

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 평균 오차제곱, ❷ 불편성, ❸ 일치성, ❹ 유효성은 바람직한 추정량에 한정되는 조건이지만, 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

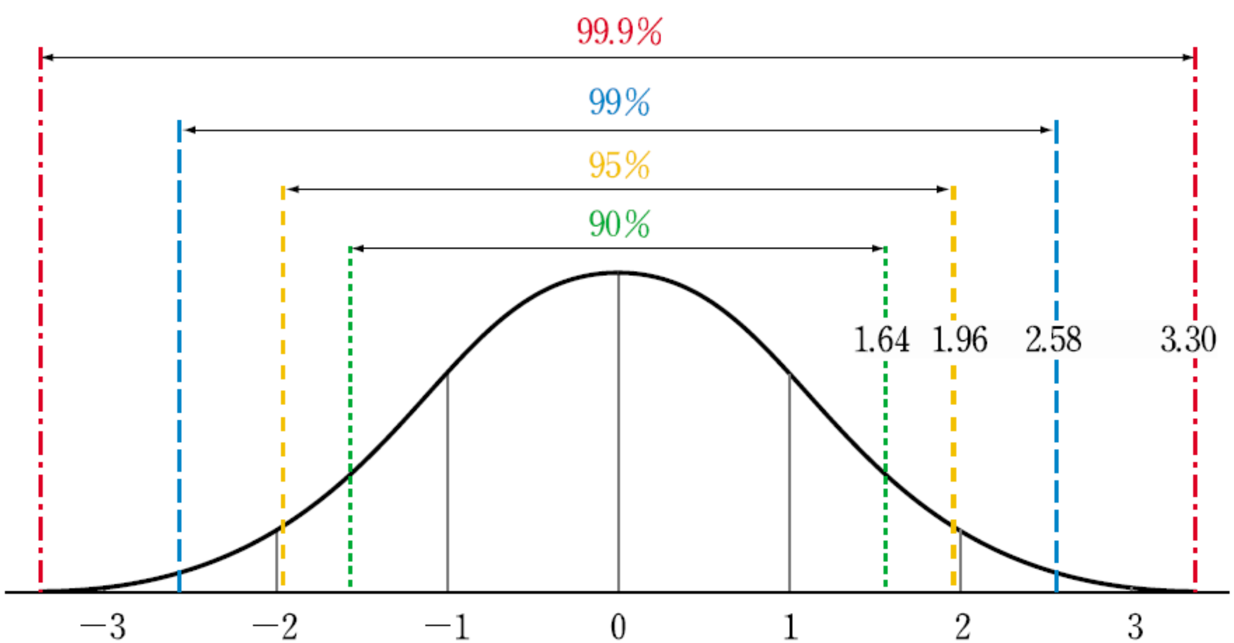
모평균(μ)을 추정할 때 표본평균을 $\overline{X}$

표준오차를 SE

표본평균 $\overline{X}$

모집단 평균에 대한 신뢰구간

$$\overline{X} - z \cdot SE \le \mu \le \overline{X} + z \cdot SE$$



구간추정

예제 6-1 구간추정

토익 점수의 평균과 표준오차가 각각 500, 100일 때, 신뢰도 90%, 95%, 99%에서 구간 추정으로 평균에 대한 모수를 추정하라.

구간추정 (예제 풀이)

- **신뢰도 90%에서의 구간추정**

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

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

- **신뢰도 95%에서의 구간추정**

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

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

구간추정 (예제 풀이)

- **신뢰도 99%에서의 구간추정**

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

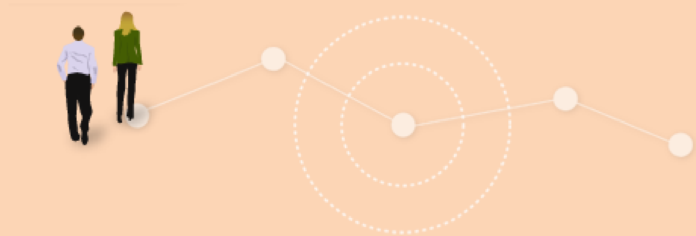
$$500 - 2.58 \cdot 100 \le \mu \le 500 + 2.58 \cdot 100$$
$$242 \le \mu \le 758$$

참고 신뢰도 100%

신뢰도를 100%로 하면 조사 결과가 모두 맞는 것이므로 최상의 결과가 될 것이다. 그러나 신뢰도 90%, 95%, 99%에서의 모평균의 구간추정 결과에서 보는 바와 같이 신뢰도가 올라갈수록 구간은 점점 넓어진다. z값이 올라가기 때문에 그런 결과가 발생하는데, 100%에 해당하는 z값은 $z = \infty$이다. 이는 $-\infty \le \mu \le \infty$의 값을 갖는다는 의미로, 틀릴 확률은 당연히 0%가 되지만 의미가 전혀 없는 결과이다. 때문에 약간의 신뢰도를 포기하고서라도 유의미한 구간을 확인하는 게 더 유용할 수 있다. 따라서 조사자가 조사 대상에 따라 신뢰수준을 정하고 구간을 추정하는 것이다.

02 모평균의 구간추정

:: Keywords z분포 | t분포 | 표본의 크기 결정



모평균의 신뢰구간 (모집단의 표준편차를 아는 경우)

■ 모집단의 표준편차를 아는 경우

모집단의 정보를 아는 경우는 거의 없음

→ 표준편차를 알고 있다고 가정하는 이유는

표준편차를 모르는 경우를 학습할 때 도움이 되기 때문

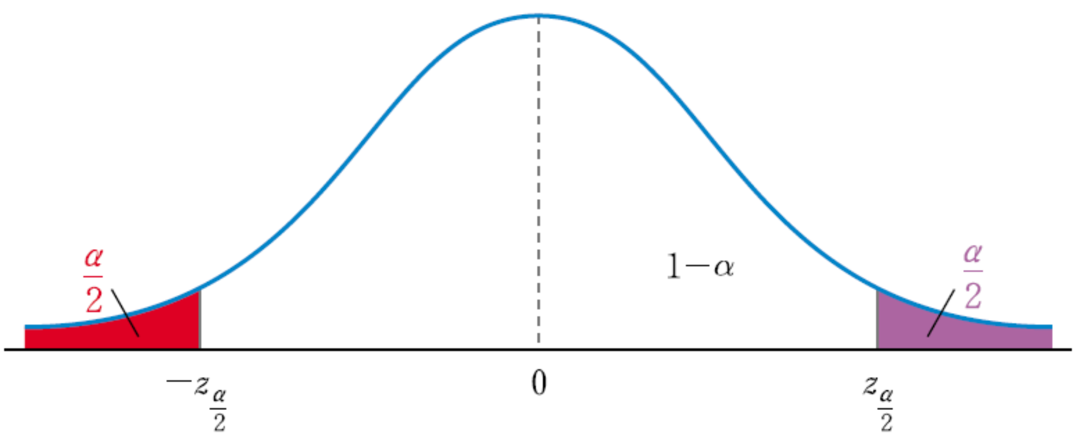
통계학에서는 90%, 95%, 99%, 99.9%에 해당하는 신뢰구간을 많이 이용

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

신뢰구간은 $1-\alpha$로 표현

모평균의 신뢰구간 (모집단의 표준편차를 아는 경우)

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



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

$$\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

모평균의 신뢰구간 (모집단의 표준편차를 아는 경우)

참고 **신뢰구간 $1-\alpha$의 유도 과정**

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

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

$$-z_{\frac{\alpha}{2}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\frac{\alpha}{2}}$$

$$-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu \le z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

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

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

모평균의 신뢰구간 (모집단의 표준편차를 아는 경우)

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

A회사에서 생산하는 전구의 평균 수명을 확인하고자 한다. 무작위로 200개의 표본을 추출하여 수명을 측정했더니 평균 수명이 30,000시간, 모표준편차는 500시간이었다. 이때 95%의 신뢰수준으로 모평균의 신뢰구간을 추정하라.

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.[4]

■ 표준편차

3장에서 중심경향도를 파악하고 분포를 나타내는 산포도를 확인하면서 분산과 표준편차를 살펴보았다. 표준편차는 표본평균으로부터 표본들의 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 표준편차이다. 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 개라는 것을 의미한다. 단순히 1개의 평균으로 표본을 추출하는 것보다는 2개 이상의 표본평균으로 모수를 추정한다면 더 정확한 추정될 수 있다. 표본의 추출 횟수를 최대한 늘려서 $n \rightarrow \infty$로 할 수 있다면, 이들의 평균은 모수와 일치하게 될 것이다. 다시 말해, 모집단에서 m개의 표본을 추출해서 m개$(\overline{x}_1, \overline{x}_2, \overline{x}_3, \cdots, \overline{x}_m)$의 평균을 구하면 1개의 표본을 통한 모평균의 추정보다 모수에 더 근접한 추정이 가능하다. 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 모수의 추정이 정확하게 이루어졌음을 판단할 수 있다. 평균의 표준오차는 다음과 같이 구한다.

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

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

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)

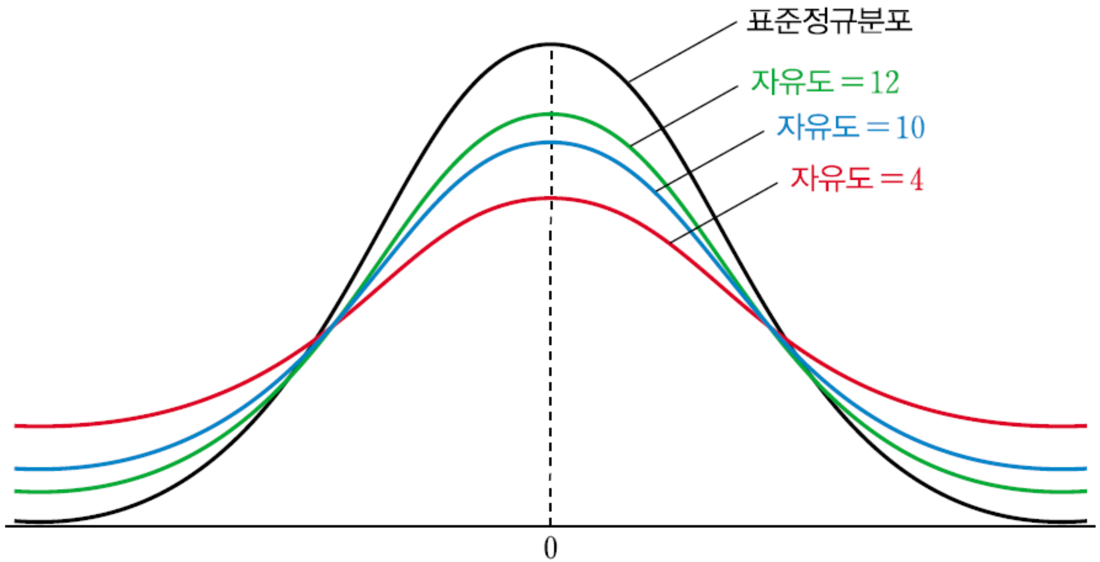
표본의 표준편차를 이용한 신뢰구간은

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 표준편차를 모르는 경우 모평균의 신뢰구간

준비파일 | 6장_신뢰구간2.xlsx

다음은 같은 반 학생 12명의 신장을 조사한 자료이다. 학생들 신장에 대해 95%의 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

참고 t분포를 활용한 t검정

t분포를 이용하여 t검정을 실시하는데, t검정은 비교할 수 있는 두 집단 내의 변화량을 이용해 서로 통계적으로 유의한 차이가 있는지를 검증할 수 있다.

$$t = \frac{(\text{두 집단의 평균 차})}{(\text{두 집단 간의 변화량})} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

이때 $s_{\overline{x}}$는 $\dfrac{s}{\sqrt{n}}$로 표준오차를 의미한다. t분포를 이용한 검정은 한 개의 표본을 대상으로 하는 일표본 t검정 외에도 대응표본 t검정, 독립표본 t검정으로 구분되는데, 이에 대한 자세한 내용은 8장에서 다루기로 한다.

표본의 크기 결정

표본의 크기가 커진다는 것 → 통계량에 대한 신뢰도가 높아지는 것

표본의 크기가 작아진다는 것 → 모집단을 대표할만한 대표성과 신뢰도 ↓

그렇다면, 적절한 표본의 크기는?

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

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

표본의 크기 결정

참고 허용오차의 유도 과정

$z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$를 허용오차 d라 할 수 있는데, $\overline{x}$는 허용오차가 100(1−α)% 확률의 신뢰구간에 포함된다.
이는 $P(|\overline{x} - \mu| \le d) = 1 - \alpha$로 표현할 수 있다. 따라서 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 **표본의 크기 결정** 준비파일 | 6장_표본개수.xlsx

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

(a) 최소 몇 개의 표본을 검사해야 하는지 표본의 개수를 구하라.

(단, 신뢰수준 99%, 허용오차 ±100ml, 표준편차 150ml)

(b) 알려진 표준편차가 없어서 10번의 표본을 추출해서 (표준편차) = 170ml를 계산했다. 신뢰수준 99%, 허용오차 ±100ml일 때, 몇 개의 표본으로 조사해야 하는지 구하라.