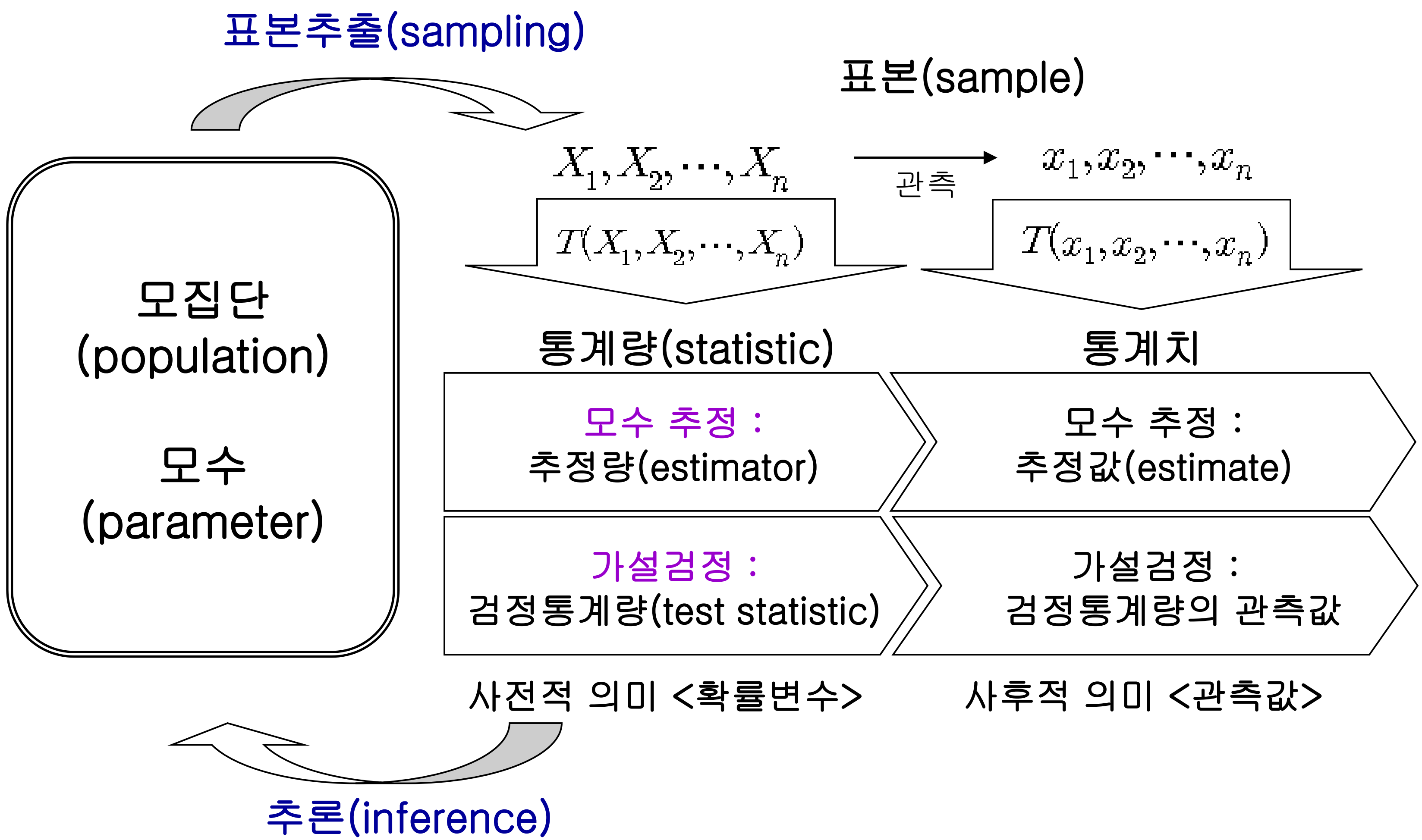
# 제 8 장 통계적 추정(모평균&모비율)

담당교수 : 김 덕 기

toby123@cbnu.ac.kr

# 통계적 추론(Inference)과정



# 통계적 추론의 예.

[예제] 특정 음식의 콜레스테롤 함유량을 알아보기 위해 50개의 표본을 추출하여 다음과 같은 자료를 얻었다.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 11.70 | 11.02 | 11.24 | 11.12 | 12.23 | 10.32 | 10.33 | 10.89 | 11.88 | 10.72 |
| 10.86 | 11.05 | 11.23 | 9.67 | 10.88 | 11.36 | 10.65 | 11.33 | 12.00 | 10.71 |
| 10.90 | 10.74 | 11.42 | 10.03 | 10.35 | 10.31 | 11.85 | 10.88 | 10.97 | 10.77 |
| 11.06 | 11.68 | 10.82 | 10.16 | 9.89 | 10.66 | 10.94 | 11.14 | 10.28 | 10.35 |
| 10.87 | 11.14 | 10.79 | 10.65 | 11.07 | 11.43 | 10.98 | 10.92 | 11.20 | 11.49 |

이 음식의 콜레스테롤 평균함유량 ( $\mu$ )에 대해 다음과 같은 질문을 생각해보자.

(1) $\mu$ 의 참 값을 어떻게 추측할 수 있는가? ➔ 점 추정의 문제.

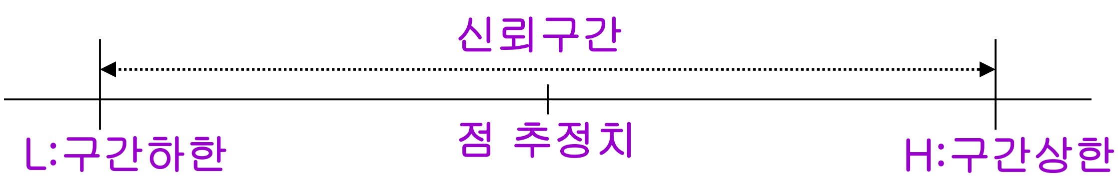
(2) $\mu$ 의 참 값이 포함되리라고 기대할 수 있는 범위를 추측할 수 있는가? ➔ 구간추정의 문제.

(3) 만일 이 음식의 평균 콜레스테롤 함유량이 10보다 낮다는 주장이 옳다고 할 수 있는가?
➔ 9장 검정의 문제.

# 점 추정 – 구간추정

| 모집단 | | 통계량 | 관측 | 통계치 |
|---|---|---|---|---|
| 모수 $\theta$ | n(=50)개의 표본추출 → / ← 추정 | $\hat{\theta} = \frac{\sum X_i}{n} = \overline{X}$ | $x_1 = 11.70$ ⋮ $x_{50} = 11.49$ | $\hat{\theta} = \overline{x} = 10.94$ |
| | | 추정량 | | 추정치 |

> 점 추정은 특정한 값으로 모수를 추정하는데 비해 구간추정은 미지의 모수의 참 값이 포함될 것으로 기대하는 범위를 구하여 추정하는 방법.




- 신뢰도(신뢰수준) : 모수가 신뢰구간에 포함되어 있을 가능성(확률).
- 90%, 95%, 99%의 신뢰도(신뢰수준)을 갖는 신뢰구간.
- 모수가 신뢰구간 안에 포함되지 않을 확률 : 오차율( $\alpha$ : 0.1, 0.05, 0.01).

# 추정량-신뢰구간-신뢰수준

표 모수와 추정량

| | 모집단(모수 $\theta$) | 표본(추정량 $\hat{\theta}$) |
|---|---|---|
| 평균 | $\mu$ | $\overline{X}$ |
| 분산 | $\sigma^2$ | $S^2$ |
| 표준편차 | $\sigma$ | $S$ |
| 비율 | $p$ | $\hat{p}$ |

## 구간추정(신뢰구간)

구간추정이란 모수를 추정하기 위하여 점추정량에 오차의 한계를 고려하여 하한 (lower limit)과 상한 (upper limit)을 포함한 구간을 제시하는 것을 의미한다.

## 신뢰수준 $1-\alpha$

신뢰수준 $1-\alpha$란 모수의 참값이 주어진 신뢰구간 $(L, U)$ 내에 포함되는 것을 얼마나 신뢰할 수 있는가를 나타내는 정도를 의미하고, 자주 사용하는 신뢰수준의 값에는 0.90, 0.95, 0.99가 있다.

# 모평균에 대한 구간추정-1.

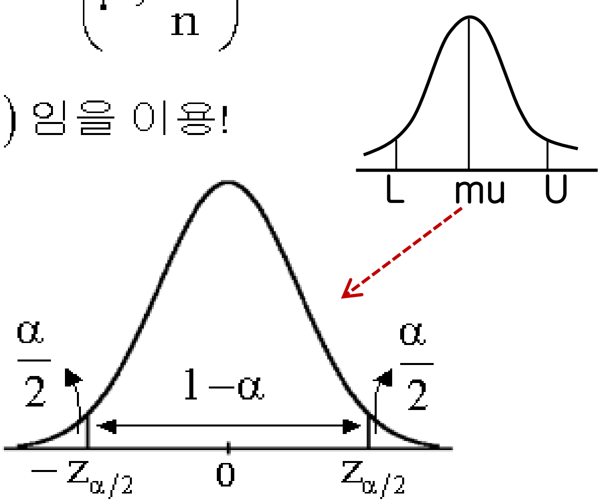
정규 모집단의 경우 : σ를 알 때

$X_1, X_2,$ ☹ $\ldots X_n \overset{r.s}{\sim} N(\mu,\sigma^2)$ ➡ $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

➡ $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ 임을 이용!

$$P\left\{-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right\} = 1-\alpha$$

$$\Leftrightarrow P\left\{\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$




μ에 대한 $100(1-\alpha)\%$ 신뢰구간

$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ 또는 $\overline{X} \pm \boxed{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$

오차한계

# 모평균에 대한 구간추정-2.

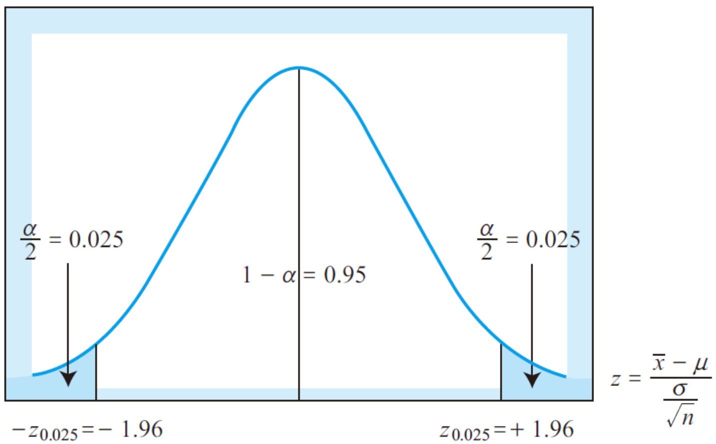



그림 신뢰수준 $1-\alpha$가 0.95일 때 $Z$의 분포에서 $\pm z_{0.025}$의 위치

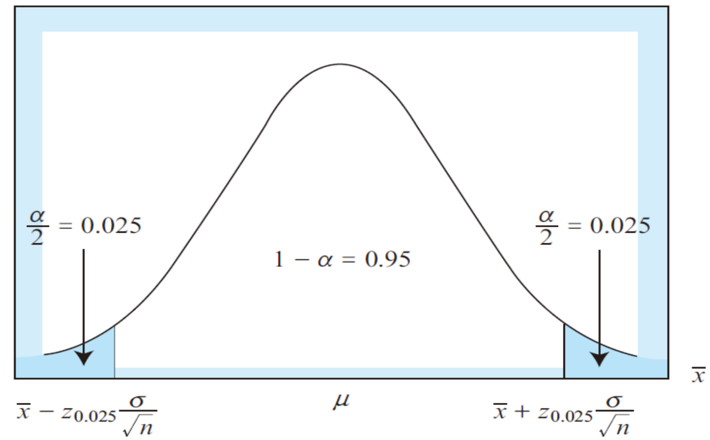



그림 신뢰수준 $1-\alpha$가 0.95일 때 $\bar{X}$의 분포에서 $\mu$의 신뢰한계

**모평균 $\mu$의 신뢰구간: 대표본의 경우**

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\quad \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

| | |
|---|---|
| $1-\alpha$ | : 신뢰수준, |
| $\bar{X}$ | : 모평균 $\mu$의 점추정량, |
| $\frac{\sigma}{\sqrt{n}}$ | : 점추정량 $\bar{X}$의 표준편차, |
| $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ | : 신뢰수준 $1-\alpha$일 때 오차의 한계, |
| $\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ | : 모평균 $\mu$의 신뢰하한, |
| $\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ | : 모평균 $\mu$의 신뢰상한. |

# 90% 신뢰수준 이란 ? ~여러 신뢰구간들

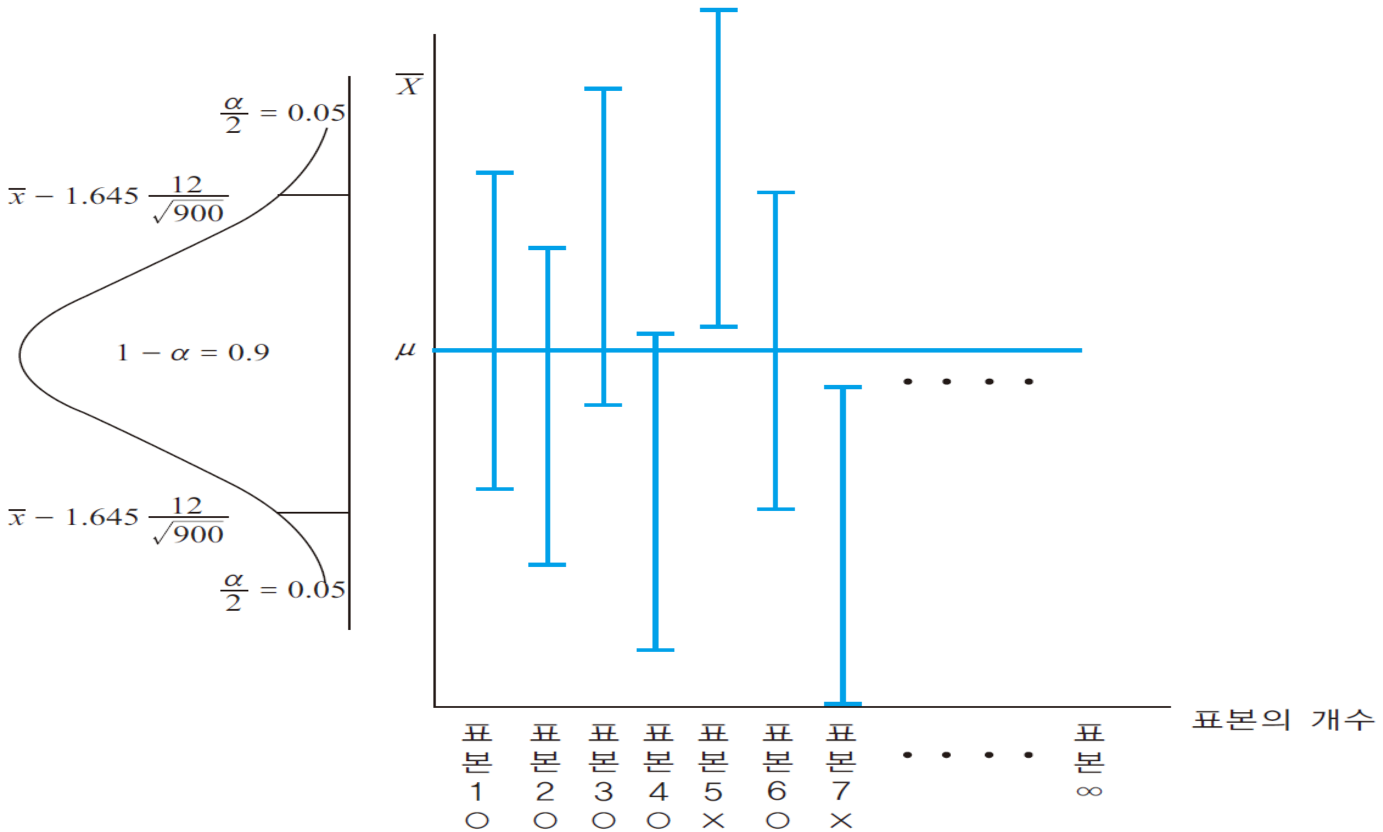



그림　표본들로부터 계산된 $\mu$의 90% 신뢰구간들

# 신뢰구간의 폭을 줄이는 방법

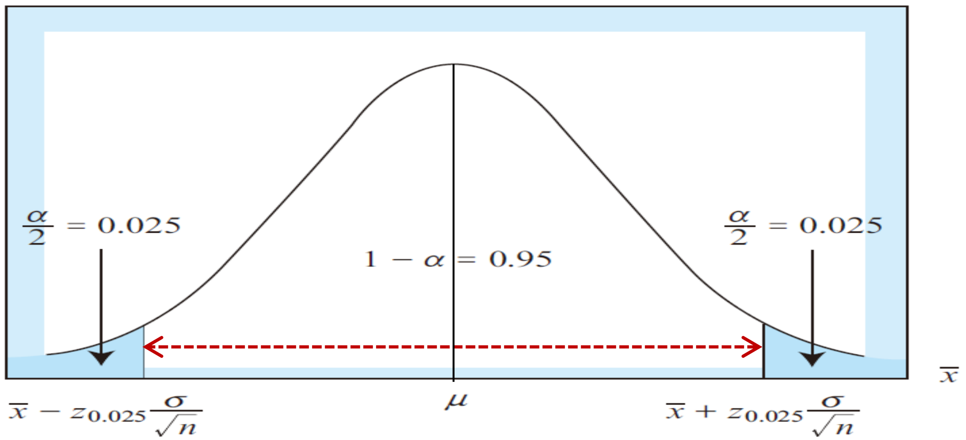



그림 신뢰수준 $1-\alpha$가 0.95일 때 $\bar{X}$의 분포에서 $\mu$의 신뢰한계

표 신뢰구간의 폭을 줄이기 위한 방법

| 방법 | 신뢰수준 $1-\alpha$ | $z_{\alpha/2}$값 | 표본의 크기 $n$ | 표준오차 $\frac{\sigma}{\sqrt{n}}$ | 오차한계 $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ | 신뢰구간 폭 $U-L$ |
|---|---|---|---|---|---|---|
| 1 | 감소 | 감소 | 고정 | 고정 | 감소 | 감소 |
| 2 | 고정 | 고정 | 증가 | 감소 | 감소 | 감소 |

# T-분포 활용 : 소 표본, 표준편차 모름

$X_1, X_2, \cdots, X_n$이 정규모집단 $N(\mu, \sigma^2)$으로부터의 확률표본일 때, 표본평균 $\overline{X}$에 대하여

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ 즉, } Z = \frac{\overline{X} - \mu_{\overline{x}}}{\sigma / \sqrt{n}}$$

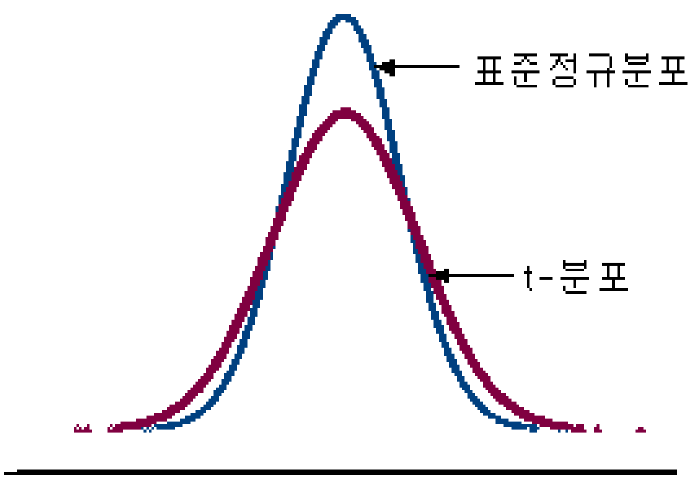
표본의 크기가 작고, 모집단의 $\sigma$을 모르는 경우 :

$\sigma$ => 표본표준편차 $S = \sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 / (n-1)}$, Z => $t = \dfrac{\overline{X} - \mu_{\overline{x}}}{S / \sqrt{n}}$

$$Z \sim N(0,1) \Rightarrow t \sim t(n-1)$$

# 정규분포-T분포

아래 그림에서 보는 바와 같이 표준정규분포와 비슷하게 평균이 0이고 분포의 모양이 평균을 중심으로 좌우가 대칭이다. 그러나 t-분포는 표준정규분포보다 평균 주위의 높이가 낮고 양쪽 꼬리 근처가 더 두꺼운 모양을 갖고 있다.



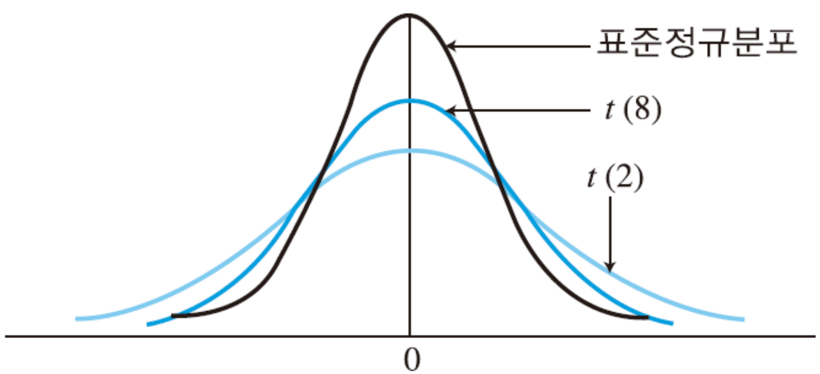




그림 표준정규분포와 자유도가 2와 8인 *t*-분포

표본의 크기가 (n>30) 커지면 t-분포도 정규분포에 근사 한다.

# 모집단분산을 모르고 소표본인 경우

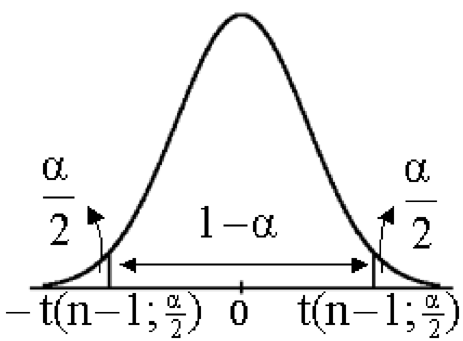
- **정규 모집단의 경우 : $\sigma$를 모를 때**

$$X_1, X_2, \cdots, X_n \overset{r.s}{\sim} N(\mu, \sigma^2) \;\Rightarrow\; \hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\Rightarrow\; T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1) \text{ 임을 이용!}$$

$$P\left\{-t(n-1;\tfrac{\alpha}{2}) \le \frac{\overline{X}-\mu}{S/\sqrt{n}} \le t(n-1;\tfrac{\alpha}{2})\right\} = 1-\alpha$$

$$\Leftrightarrow P\left\{\overline{X} - t(n-1;\tfrac{\alpha}{2})\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t(n-1;\tfrac{\alpha}{2})\frac{S}{\sqrt{n}}\right\} = 1-\alpha$$




$\mu$에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\left(\overline{X} - t(n-1;\tfrac{\alpha}{2})\frac{S}{\sqrt{n}},\ \overline{X} + t(n-1;\tfrac{\alpha}{2})\frac{S}{\sqrt{n}}\right) \text{ 또는 } \overline{X} \pm t(n-1;\tfrac{\alpha}{2})\frac{S}{\sqrt{n}}$$

# (…예제)

도로의 차선을 긋는데 사용하는 페인트의 내구성을 검사하려고 한다. 8개 지역을 선정하여 차선을 긋고 차선의 퇴색상태를 조사한 결과 차량의 통과횟수가 142,600, 167,800, 136,500, 108,300, 126,400, 133,700, 162,000, 149,000 회를 넘어서면서 차선이 퇴색하기 시작하였다. 이 페인트의 내구성을 차량의 평균통과횟수의 측면에서 95%의 신뢰구간으로 추정하라.

(단, 차량 통과횟수의 분포는 정규분포라고 가정한다.)

$$\bar{x}: \frac{\sum x_i}{n} = 140{,}800 \ , \quad s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = 19{,}200$$

$$P\left(140{,}800 - 2.365\frac{19{,}200}{\sqrt{8}} \leq \mu \leq 140{,}800 + 2.365\frac{19{,}200}{\sqrt{8}}\right) = 0.95$$

$$\therefore \ 124{,}700 \leq \mu \leq 156{,}900$$

$$\boxed{\alpha = 0.05 : \ t_{\alpha/2}(n-1) = t_{0.025}(7) = 2.365}$$

# 비 정규모집단의 경우

- **비정규 모집단의 경우**

$X_1, X_2, \cdots, X_n \overset{r.s}{\sim} (\mu, \sigma^2)$ ➡ n 이 충분히 큰 경우, $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\bullet}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$

➡ $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \overset{\bullet}{\sim} N(0,1)$ 임을 이용! ( $\sigma$를 알 때)

➡ $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \overset{\bullet}{\sim} N(0,1)$ 임을 이용! ( $\sigma$를 모를 때)

$\mu$에 대한 $100(1-\alpha)\%$근사 신뢰구간

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$ ( $\sigma$를 알 때)

$$\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$ ( $\sigma$를 모를 때)

# 모비율에 대한 신뢰구간추정.

| 모비율의 구간추정 |
| --- |
| |
| 모비율에 대한 $100(1-\alpha)\%$ 신뢰구간은 $$\left(\bar{p}-z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}},\ \bar{p}+z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right)$$ |

K조사연구소에서는 한국의 중소기업체 경영자 202명을 대상으로 한국경제의 미래에 관한 전망을 조사하였다. 그 결과 105명이 낙관적인 전망을 하는 것으로 나타났다. 한국경제에 대해 낙관적인 전망을 하는 중소기업 경영자의 비율은 어느 정도인가를 95%신뢰도로 추정하라.

$\bar{p}=105/202=0.5198$ , $Z_{\alpha/2}=1.96$ , $n=202$

$100(1-\alpha)\%:\alpha=0.05\ \longrightarrow$ 95% 신뢰구간은 다음과 같다.

$$\left(\bar{p}-z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}},\ \bar{p}+z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right)$$

$$\longrightarrow 0.5198\pm1.96\sqrt{\frac{0.5198(1-0.5198)}{202}}\quad :\ 0.4512\leq p\leq0.5884$$

# 모평균 추정에서 표본크기의 결정

μ에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad \text{또는} \quad \overline{X}\pm\boxed{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

오차한계

➔ E는 신뢰구간의 폭의 반을 나타내고, 신뢰구간의 폭이 지나치게 크면 의사결정을 위한 정보의 가치가 줄어 듬 ➔ 신뢰구간의 폭을 가능한 한 줄여 정보가치를 높임 ➔ n이 증가하거나 z가 감소하면 신뢰구간은 줄어 든다. ➔ 특정수준의 신뢰도하에서 신뢰구간의 정해진 수준만큼 유지하기 위해 표본의 크기 n을 결정.

**모평균의 추정에서 표본크기의 결정**

오차의 한계를 $E$ 로 하려면, 신뢰수준 $1-\alpha$를 만족해 주도록 하는 표본의 크기는

$$n=\frac{(z_{\alpha/2})^2\cdot\sigma^2}{E^2} \quad \text{또는} \quad n=\frac{(t_{\alpha/2}(n-1))^2\cdot S^2}{E^2}$$

분산을 모를 때

분산을 알 때

# 모비율 추정에서 표본크기의 결정

P에 대한 (1−α)100% 신뢰구간

오차한계

$$\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \text{ 또는 } \hat{p}\pm\boxed{z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

$$E \geq z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

위 식을 $n$에 관해 풀면 $n \geq \dfrac{(z_{\alpha/2})^2 \cdot p(1-p)}{E^2}$

① $p$에 대한 지식이 없는 경우 : $n \geq \dfrac{(z_{\alpha/2})^2}{4\,E^2}$

② 과거의 조사에 의해 표본비율을 알고 있는 경우 : $n \geq \dfrac{(z_{\alpha/2})^2 \cdot p^*(1-p^*)}{E^2}$

$P(1-P) = -P^2 + P \rightarrow$ 볼록함수(미분 $=0$) : $P = 1/2$일때 최대값