제 9 장 정규모집단에서의 추론

담당교수 : 김 덕 기

toby123@cbnu.ac.kr

모집단의 정규분포에 대한 가정 확인

- 자료를 분석할 때 모집단이 정규분포를 따른다고 가정하는 경우가 많다.

- 관측한 자료가 정규분포를 따르는지 확인하는 방법은?

  - 자료에 대한 히스토그램이 정규분포에 가까운지 확인
    → 주관적 판단. 그리는 방법에 따라 해석이 달라질 수 있다.

  - $X \sim N(\mu, \sigma^2)$일 때
    $P(\mu - \sigma \le X \le \mu + \sigma) = 0.6827$
    $P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.9545$
    $P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.9973$
    이므로, 관측값 중에서
    $(\bar{x} - s, \bar{x} + s), (\bar{x} - 2s, \bar{x} + 2s), (\bar{x} - 3s, \bar{x} + 3s)$
    에 속하는 비율과 위의 확률이 유사한지 확인

정규확률그림(normal probability plot)

- 자료의 백분위수와 정규분포의 백분위수를 그림을 통해 비교하여 자료가 정규분포를 따르는지 확인

- n개의 자료: $x_1, x_2, \ldots, x_n$
  ⇒ $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$: 크기 순으로 배열된 자료

- $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$은 0과 1 사이의 확률을 균등하게 $(n+1)$ 등분하는 백분위수들이라고 할 수 있다.

- n개의 자료가 정규분포를 따르는 모집단 $N(\mu, \sigma^2)$에서 얻어진 자료라면 모집단 $N(\mu, \sigma^2)$을 균등하게 $(n+1)$ 등분하는 n개의 점 $a_{(1)}, a_{(2)}, \ldots, a_{(n)}$과 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$은 유사할 것이다.

정규확률그림 1

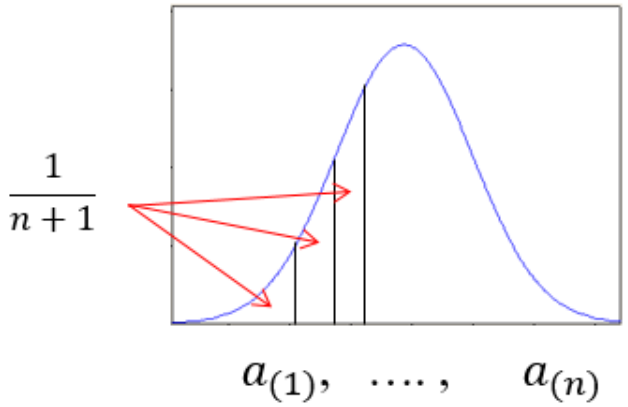
- $X \sim N(\mu, \sigma^2)$일 때

$$P(X \le a_{(1)}) = \frac{1}{n+1}$$

$$P(a_{(1)} \le X \le a_{(2)}) = \frac{1}{n+1}$$

$$P(a_{(2)} \le X \le a_{(3)}) = \frac{1}{n+1}$$

...



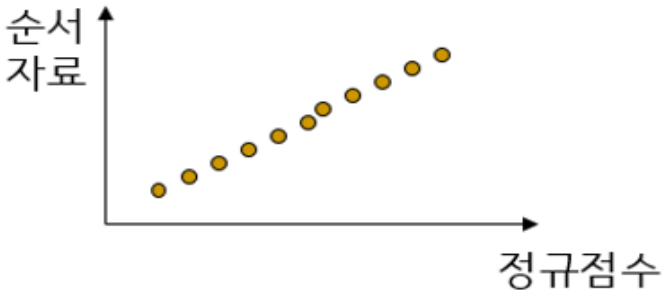
대안) 표준정규분포 $N(0,1)$을 $(n+1)$ 등분하는 n개의 점 $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$ (정규점수, normal scores)이라고 할 때

$$a_{(i)} \approx \mu + \sigma z_{(i)}$$

이 성립하므로 $(x_{(1)}, z_{(1)}), (x_{(2)}, z_{(2)}), \ldots, (x_{(n)}, z_{(n)})$을 좌표평면에 점으로 나타내면 직선에 가까워야 한다.

정규확률그림 2

$(x_{(1)}, z_{(1)}), (x_{(2)}, z_{(2)}), \ldots, (x_{(n)}, z_{(n)})$을 좌표평면에 점으로 나타낸다.



정규확률그림이 직선에 가까우면 모집단이 정규분포를 따른다고 할 수 있다.

- 자료가 정규분포를 따르지 않는다면 어떻게 할 것인가?
- 원 자료가 정규분포를 따르지 않더라도 변환을 하면 정규분포를 따를 수 있다.
- 유용한 몇 가지 변환
  - $x, x^2, \sqrt{x}, x^{1/4}, \log x, \frac{1}{x}$ 등

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

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

$X \sim (\mu, \sigma^2)$
연속형 → N 비율 → 2항 분포 (확률로 모름)
n(大) : CLT. X ~ normal
n(小) : $\overline{X}$ ~ 분포(X) 비모수적 방법 - 대안

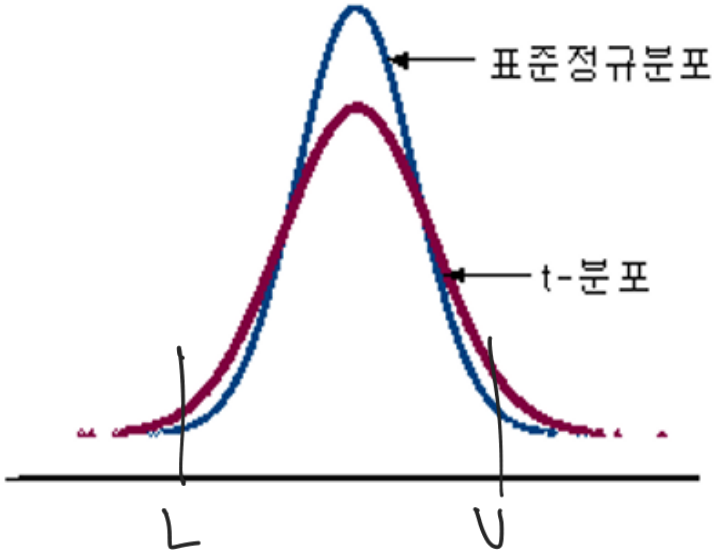
표본의 크기가 작고, 모집단의 σ을 모르는 경우 :

σ => 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, T-분포의 관계 및 특징

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.



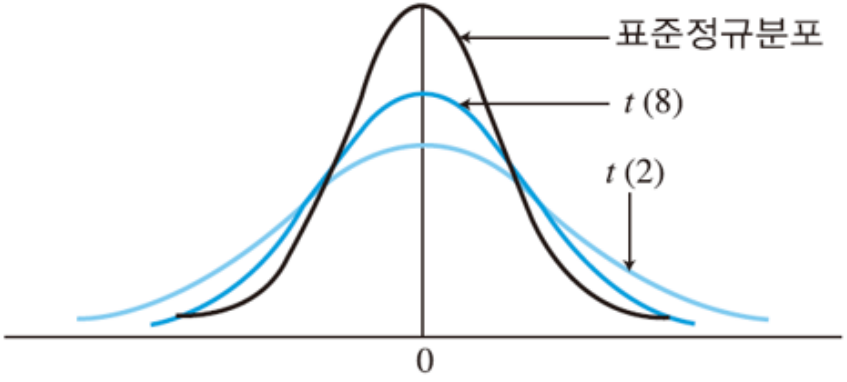


그림 : 표준정규분포와 자유도가 2와 8인 t-분포

표본의 크기가 (n>30) 커지면 t-분포도 정규분포에 근사 한다.

T-분포를 이용한 평균의 신뢰구간 추정

정규 모집단의 경우 : σ를 모를 때

[기] 정의 : 추정량 ± 표본오차

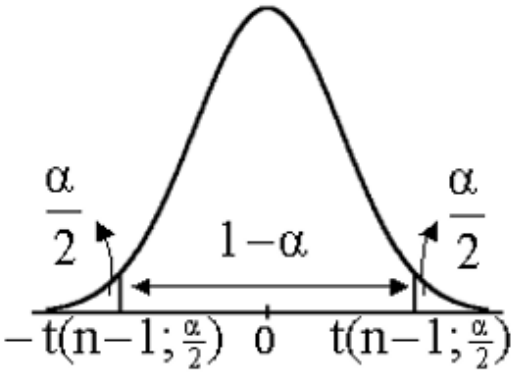
$$X_1, X_2, \cdots, X_n \overset{r.s}{\sim} N(\mu, \sigma^2) \Rightarrow \hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\mu \rightarrow \bar{x} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$

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

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

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



μ에 대한 $100(1-\alpha)\%$ 신뢰구간

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

정규분포와 T분포의 신뢰구간 비교

- μ에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$\overline{X} \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}$$

- 모표준편차 σ가 알려진 경우 μ에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- $t-$분포는 표준정규분포에 비해 꼬리가 두꺼우므로 $t_{\alpha/2}(n-1) > z_{\alpha/2}$가 성립한다. 따라서 σ가 알려진 경우에 비해 신뢰구간의 길이가 길어진다.

T-분포를 이용한 신뢰구간 추정 : example

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.)

$X \sim$ normal

$n=8$ (小), σ^2 모름

$\Rightarrow$ t분포

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

$$\bar{X} = t_{\alpha/2} \frac{S}{\sqrt{n}}$$

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

$$\therefore \; 124{,}700 \leqq \mu \leqq 156{,}900$$

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

비정규 모집단에서 신뢰구간 추정

- 비정규 모집단의 경우

CLT

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

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

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

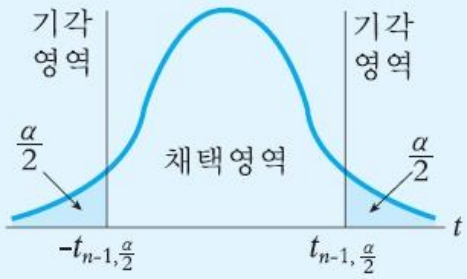
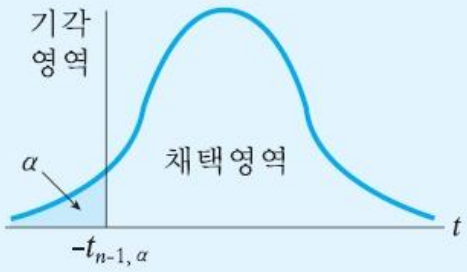
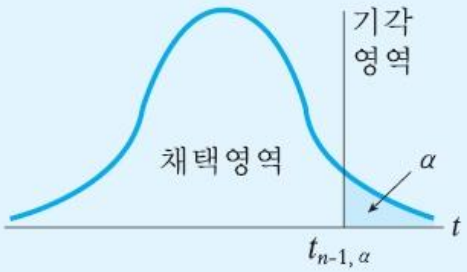
μ에 대한 $100(1-\alpha)\%$ 근사 신뢰구간

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

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

모평균에 대한 가설검정: t −검정

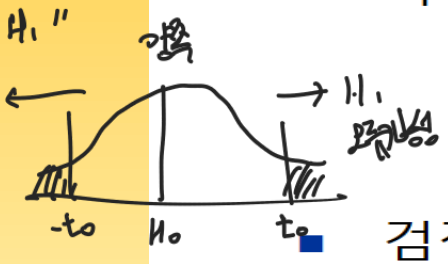
검정통계량 : $t_C = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

양측검정	좌측검정	우측검정
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$
		
만일 $t_c < -t_{n-1,\frac{\alpha}{2}}$ 또는 $t_c > t_{n-1,\frac{\alpha}{2}}$ 이면 H_0을 기각	만일 $t_c < -t_{n-1,\alpha}$ 이면 H_0을 기각	만일 $t_c > t_{n-1,\alpha}$ 이면 H_0을 기각

모든 경우에 p값$<\alpha$이면 H_0을 기각

T-분포에서 유의확률(p-value)

- t -검정에서 p-값
  - 가설:



$$H_0: \mu = \mu_0 \quad vs \quad \begin{cases} (i)\ H_1: \mu < \mu_0 \\ (ii)\ H_1: \mu > \mu_0 \\ (iii)\ H_1: \mu \neq \mu_0 \end{cases}$$

  - 검정통계량:

$$t = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S}$$

  - 주어진 관측값에 대한 검정통계량의 값이 $t = t_0$일 때

$p = 2 * P[H_1 \mid H_0]$

$= 2 * P[T \geq t_0 \mid H_0]$

$$p-\text{값} = \begin{cases} (i) & P(t \leq t_0) \\ (ii) & P(t \geq t_0) \\ (iii) & P(|t| \geq |t_0|) \end{cases}$$

단, $t \sim t(n-1)$

→ 대립 가설이 양측 검정인 경우 p-value = $2*P(Z \leq z)$ $or.$ $2 * P(Z \geq z)$

Test ① 기각역 판정
② p-value 판정
③ CI 판정

T-분포 평균검정 : example

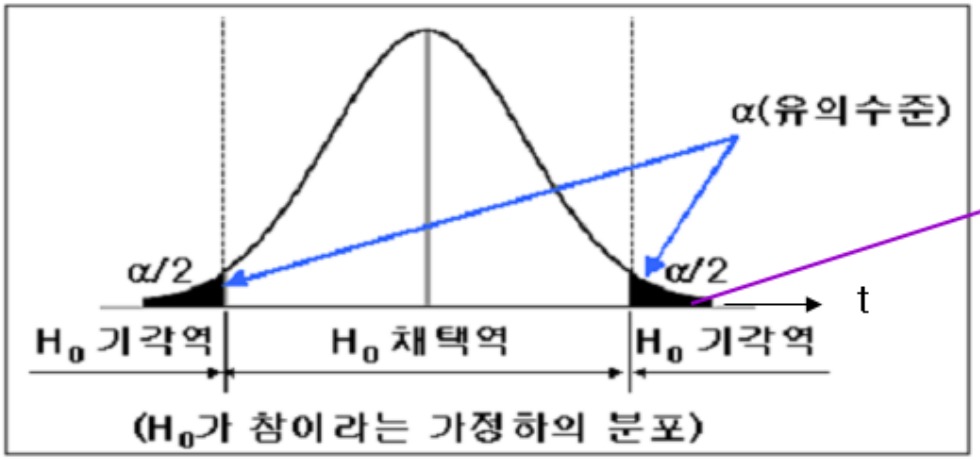
모집단의 분산을 모르는 경우 : (양쪽검정)

[예제]길이가 30cm인 제도용 자를 16개 표본 추출하여 길이를 측정한 결과 평균이 30.1cm, 표준편차는 0.04cm 였다. 제도용 자의 평균길이가 30cm라 할 수 있는지를 1%의 유의수준으로 검정하라. (정규분포를 가정)

• 단일 모평균에 대한 가설검정

	귀무가설	대립가설		(H_0하에서의) 검정통계량	유의수준 α에 대한 기각역
(a)	$H_0: \mu = \mu_0$	$H_1: \mu > \mu_0$	우측검정 ┐ 단측검정	$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$Z \geq z_\alpha$ 또는 $T \geq t_{(n-1,\alpha)}$
(b)		$H_1: \mu < \mu_0$	좌측검정 ┘	또는	$Z \leq -z_\alpha$ 또는 $T \leq -t_{(n-1,\alpha)}$
(c)		$H_1: \mu \neq \mu_0$	양측검정	$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\lvert Z \rvert \geq z_{\alpha/2}$ 또는 $\lvert T \rvert \geq t_{(n-1,\alpha/2)}$

T-분포 평균검정 : 3가지 판정방법



$|t| > t_{\alpha/2,n-1} = t_{0.005,15} = 2.947$

CI : $\bar{X} \pm t(n-1;\frac{\alpha}{2})\frac{S}{\sqrt{n}} \in H\_0$

P-value=P(t >=10)*2

① 가설설정 : $H_0 : \mu = 30$, $H_A : \mu \neq 30$

② 유의수준 : $\alpha = 0.01$

③ 검정통계량 : $\sigma^2(Unknown)$, 정규분포가정-> 검정통계량(t-통계량)

④ 유의수준 α의 기각역(임계치) : $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$ (양쪽검정)

⑤ 검정통계량 : $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{30.1-30}{0.04/\sqrt{16}} = 10 > t_{\alpha/2,n-1} = 2.947$

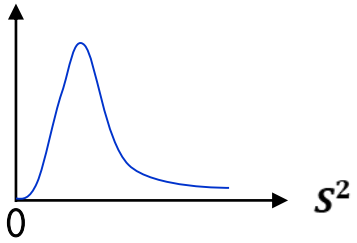
⑥ 귀무가설을 기각한다. 즉, 제도용 자의 평균길이가 30cm라 할 수 없다.

카이제곱분포(chi-square distribution)

- [정리] $X_1, X_2, \ldots, X_n$이 정규 모집단 $N(\mu, \sigma^2)$에서 임의추출한 표본일 때

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$: 자유도가 $(n-1)$인 카이제곱분포

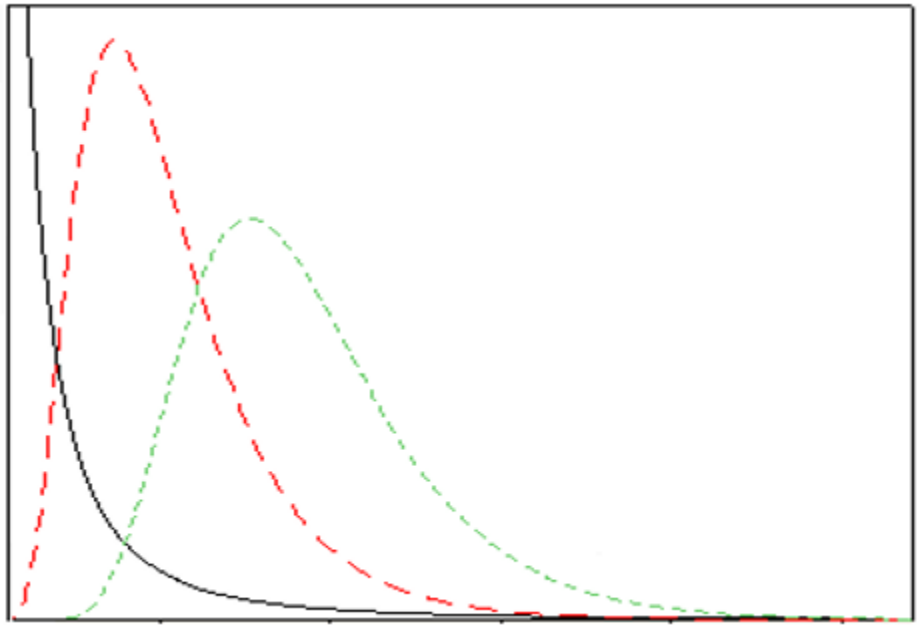
N개 모집단 $X_1, X_2, \ldots, X_N$ ➡ n개 표본을 m번 반복적으로 뽑음 $S_1^2, S_2^2 \ldots, S_m^2$



$\sum_{i=1}^{n} Z_i^2 \approx x^2(n-1), S^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$ ➔ 카이제곱 : $\frac{(n-1)S^2}{\sigma^2}$

- t −분포와 카이제곱분포의 관계
- $\frac{\sqrt{n}(\bar{X}-\mu)}{S} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} / \frac{S}{\sigma} \sim Z / \sqrt{\frac{\chi^2(n-1)}{n-1}} \sim t(n-1)$

카이제곱 분포의 특징



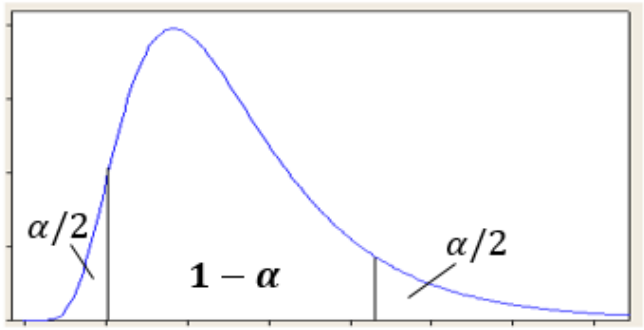
θ < N ~ ↑ / T ~ ↓

θ < B ~ ↓ / Z ~ ↑

$\theta = \sigma^2$ < χ^2 ~ 오른쪽 꼬리 / F ~ 왼쪽 이상

자유도가 커질수록 정규분포의 대칭성에 가까워진다.

모 표준편차의 신뢰구간 추정



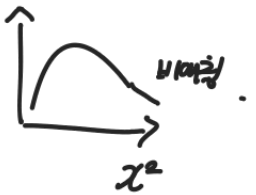
- σ에 대한 $100(1-\alpha)\%$ 신뢰구간:

$$\left(S\sqrt{\frac{(n-1)}{\chi^2{}_{\alpha/2}(n-1)}},\quad S\sqrt{\frac{(n-1)}{\chi^2{}_{1-\alpha/2}(n-1)}} \right)$$

- $P\left(\chi^2{}_{1-\alpha/2}(n-1) \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2{}_{\alpha/2}(n-1)\right) = 1-\alpha$

$$\rightarrow P\left(\frac{(n-1)S^2}{\chi^2{}_{\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2{}_{1-\alpha/2}(n-1)}\right) = 1-\alpha$$

- σ^2에 대한 $100(1-\alpha)\%$ 신뢰구간:



$$\left(\frac{(n-1)S^2}{\chi^2{}_{\alpha/2}(n-1)},\quad \frac{(n-1)S^2}{\chi^2{}_{1-\alpha/2}(n-1)}\right)$$

표준편차에 대한 가설검정

- 가설:

$$H_0: \sigma = \sigma_0 \quad vs \quad \begin{cases} (i)\ H_1: \sigma < \sigma_0 \\ (ii)\ H_1: \sigma > \sigma_0 \\ (iii)\ H_1: \sigma \neq \sigma_0 \end{cases}$$

- 검정통계량:

$$\chi^2 = \frac{(n-1)S^2}{{\sigma_0}^2} \sim \chi^2(n-1) \qquad (under\ H_0)$$

(분모에 ${\sigma_0}^2$을 사용하는 것에 주의 할 것)

- 대립가설 $H_1: \sigma < \sigma_0$이 참이라면 표본분산 S^2은 작은 값을 가질 경향이 크므로 검정통계량 χ^2도 작은 값을 가지게 될 것이다. 따라서 대립가설이 $H_1: \sigma < \sigma_0$ 일 때, 유의수준 α에서 기각역은 $\chi^2 \leq {\chi^2}_{1-\alpha}(n-1)$이다.

표준편차의 기각역 판정의 예

- 유의수준 α에서 기각역:
  - (i) $H_1: \sigma < \sigma_0$일 때, $R: \chi^2 = \frac{(n-1)S^2}{{\sigma_0}^2} \le {\chi^2}_{1-\alpha}(n-1)$
  - (ii) $H_1: \sigma > \sigma_0$일 때, $R: \chi^2 = \frac{(n-1)S^2}{{\sigma_0}^2} \ge {\chi^2}_{\alpha}(n-1)$
  - (iii) $H_1: \sigma \neq \sigma_0$일 때, $R: \chi^2 = \frac{(n-1)S^2}{{\sigma_0}^2} \le {\chi^2}_{1-\alpha/2}(n-1)$ or
    $\chi^2 = \frac{(n-1)S^2}{{\sigma_0}^2} \ge {\chi^2}_{\alpha/2}(n-1)$

예제 8) 예제 7에서 $\sigma > 0.2$인지 유의수준 0.05에서 검정

- 자료: $n = 10,\ s = 0.4$
- 가설: $H_0: \sigma = 0.2 \quad vs \quad H_1: \sigma > 0.2$
- 검정통계량:

$$\chi^2 = \frac{(n-1)S^2}{{\sigma_0}^2} = \frac{9 \times 0.4^2}{0.2^2} = 36$$

- 기각역: $\chi^2 \ge {\chi^2}_{0.05}(9) = 16.92$
- 따라서 귀무가설 $H_0: \sigma = 0.2$를 기각. 즉 σ는 0.2보다 크다고 할 수 있다.