기수 : 10 기 **이름** : 신재우

23-2 DSL 정규 세션

기초과제 1 통계적 사고

1-1 : 중심극한정리 (Central Limit Theorem) 의 정의와 그 의미를 서술하시오.

n 개의 'X' 확률변수들의 평균의 분포는 정규분포에 가까워 진다는 것이며, 이것에는 조건이 있다. 우선은 n 이 적당히 커야 하며 대부분의 경우 그 기준이 30 이다. 두번째로는 확률변수들이 모두 독립적이여야 한다는 것이다. 세번째로는 확률변수들이 모두 같은 분포를 가지고 있어야 하며 분포의 평균과 분산은 유한해야 한다. 수식으로의 표현은 아래와 같다.

$$X_1, X_2, \cdots, X_n \sim iid\ (\mu, \sigma^2) \tag{1}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \tag{2}$$

1-2 : 중심극한정리가 통계적 추론 중 "구간추정"에서 어떻게 유용한지 서술하시오.

위의 중심극한정리를 이용하자면 X 의 표본평균은 (1)의 식과 같이 정규분포를 따른다. 정규분포의 특징을 이용하면 X 의 표본평균을 표준정규분포인 Z 값으로 변형시킬 수가 있다. 수식으로의 표현은 아래와 같다.

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}) \tag{3}$$

$$\overline{X} - \mu \sim N(0, \frac{\sigma^2}{n}) \tag{4}$$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = Z \sim N(0, 1) \tag{5}$$

하지만 이것은 n 이 무한대로 극한하게 된다면 적용되기 때문에 σ^2 대신에 S^2 를 분산으로 적용해도 해당이 되지만 Z 에 근사하게 된다는 차이점이 있다. 이 이유는 X 변수에 대한 분포가 정규분포라고 더 이상 단정짓기가 어렵기 때문이다. 이 점을 이용해서 모평균인 μ 의 신뢰구간 / **구간추정** 을 구할 수가 있게 된다. 수식으로의 표현은 아래와 같으며 α 는 유의수준을 의미한다.

$$1 - \alpha \ \approx \ P(-Z_{\frac{\alpha}{2}} \ < \ \frac{\overline{X} - \mu}{S/\sqrt{n}} < \ Z_{\frac{\alpha}{2}}) \tag{6}$$

$$1 - \alpha \ \approx \ P(-Z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \ < \ \overline{X} - \mu < Z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}) \tag{7}$$

$$1 - \alpha \ \approx \ P\left(-\overline{X} - Z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \ < \ -\mu < -\overline{X} + Z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) \tag{8}$$

$$1 - \alpha \ \approx \ P\left(\overline{X} - Z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \ < \ \mu < \overline{X} + Z_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) \tag{9}$$

1-3 : 중심극한정리를 이용하여 모평균에 대한 근사신뢰구간을 만들 때, 표준오차($\sqrt{Var(\bar{X})}$) 부분의 모분산을 표본분산으로 대체할 수 있는 이유를 수식적으로 증명하시오.

궁극적으로 우리의 목표는 (5)의 식에서 모분산을 표본분산으로 대체가 가능한지를 증명하고 싶다. 우선은 표본분산인 S^2 가 모분산인 σ^2 에 확률 수렴한다는 것을 먼저 보여주고 나서 Slutsky's Theorem 를 이용해서 이것을 (5)의 식에도 적용이 가능하다는 것을 보여주겠다.

$$S^2{}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i{}^2 - X_i\bar{X}_n - X_i\bar{X}_n + \bar{X}_n{}^2)$$
$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n} X_i{}^2 - 2\bar{X}_n\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\bar{X}_n{}^2\right\}$$
$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n} X_i{}^2 - 2\bar{X}_n n\bar{X}_n + n\bar{X}_n{}^2\right\} = \frac{1}{n-1}\left\{\sum_{i=1}^{n} X_i{}^2 - 2n\bar{X}_n{}^2 + n\bar{X}_n{}^2\right\} \tag{10}$$
$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n} X_i{}^2 - n\bar{X}_n{}^2\right\} = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i{}^2 - \bar{X}_n{}^2\right) \xrightarrow{P} 1 \times (E[X_1^2] - \mu^2)$$
$$= \sigma^2$$

n 은 상수이기 때문에 다음과 같은 확률수렴을 제시할 수 있다.

$$\lim_{n\to\infty}\frac{n}{n-1} = 1 \tag{11}$$
$$\frac{n}{n-1} \xrightarrow{P} 1$$

Slutsky's Theorem 을 적용시켜서 (10)의 식에서의 확률수렴을 성립하게 만든다.

(10)에서의 확률수렴을 Slutsky's Theorem 과 함께 (5)의 식에 적용시키게 된다면 다음과 같이 수식으로 표현이 된다.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$$
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{P} \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{12}$$
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{D} N(0,1)$$

번외로 Slutsky's Theorem 이란 X_n, X, A_n, B_n 들이 확률변수이며 a와 b가 상수이면서$X_n \xrightarrow{D} X,\ A_n \xrightarrow{P} a,\ B_n \xrightarrow{P} b$ 일 때 다음과 같은 공식이 Slutsky's Theorem 를 의미한다.

$$A_n + B_n X_n \xrightarrow{D} a + bX$$

2-1 : 스튜던트 정리의 3 번째 내용을 작성 및 증명하시오.

③ $(n-1)S^2/\sigma^2$ 는 $\chi^2(n-1)$ 분포를 따른다.

우리는 $X_1, \cdots, X_n \sim^{iid} N(\mu, \sigma^2)$ 이라고 가정한다면 카이제곱분포를 다음과 같이 사용할 수 있다.

$$V = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \xrightarrow{D} \chi^2(n) \tag{13}$$

이 식에서 $X_i - \mu$ 부분을 표본분산을 나타내는 식으로 바꿔야 되며 수식으로는 다음과 같이 표현하였다.

$$\begin{aligned} V &= \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left({X_i}^2 - \mu^2\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(({X_i}^2 - \overline{X}^2) + (\overline{X}^2 - \mu^2)\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left((X_i - \overline{X})^2 + (\overline{X} - \mu)^2\right) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(\overline{X} - \mu)^2 = \frac{1}{\sigma^2} \cdot \frac{n-1}{n-1} \cdot \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + \frac{n}{\sigma^2}\left(\overline{X} - \mu\right)^2 = \frac{n-1}{\sigma^2} \cdot S^2 + \left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \xrightarrow{D} \chi^2(n) \end{aligned} \tag{14}$$

$$\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 = (Z_n)^2 \xrightarrow{D} \chi^2(1) \tag{15}$$

식 (14)에서의 마지막 부분에서 좌측 값과 우측 값 (15) 는 독립이기 때문에 마지막 부분에 mgf 를 취하게 된다면 증명이 마무리된다. 카이제곱분포의 mgf 공식은 $(1-2t)^{-n/2}$ 이며 이것을 (15)에 활용하면 $(1-2t)^{-1/2}$ 이 된다. 이것을 (14)의 마지막 부등호에 적용시킨다면 다음과 같이 수식으로 표현이 가능하다.

$$\begin{gathered} (1-2t)^{-\frac{n}{2}} = E\left[e^{t \cdot \frac{n-1}{\sigma^2} S^2 + t\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2}\right] = E\left[e^{t \cdot \frac{n-1}{\sigma^2} \cdot S^2}\right] \cdot (1-2t)^{-\frac{1}{2}} \\ (1-2t)^{-\frac{n}{2}} \cdot (1-2t)^{\frac{1}{2}} = E\left[e^{t \cdot \frac{n-1}{\sigma^2} S^2}\right] = (1-2t)^{-\frac{(n-1)}{2}} \end{gathered} \tag{16}$$

식 (16)의 우측은 $\chi^2(n-1)$ 의 mgf 형태를 띄우고 있기 때문에 $\frac{n-1}{\sigma^2} \cdot S^2$ 은 $\chi^2(n-1)$ 의 분포를 따라간다고 마무리 할 수가 있다.

2-2 : 스튜던트 정리의 4 번째 내용을 작성 및 증명하시오

④ $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t\,(n-1)$

우선은 t-분포와 카이제곱분포 그리고 표준정규분포의 관계를 식으로 보여줘야 한다. 해당 식의 카이제곱분포 부분에 2-1 의 식을 대입시키면서 식을 풀게 된다면 2-2 의 증명은 마무리된다. 수식으로는 다음과 같이 표현할 수 있다.

$$\begin{gathered} T = \frac{Z}{\sqrt{V/k}} \sim t(k) \\ T = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{n-1}{\sigma^2} \cdot S^2/(n-1)}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\frac{S}{\sigma}} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1) \end{gathered} \tag{17}$$

3-1 (a) : 귀무가설과 대립가설을 설정하시오.

Let DSL : DSL 사람들의 키, $NDSL$: DSL이 아닌 사람들의 키,
μ_{DSL} : DSL 사람들의 평균 키, μ_{NDSL} : DSL이 아닌 사람들의 평균 키
$\mathrm{D} = \mathrm{DSL} - \mathrm{NDSL}$, $\mu_D = \mu_{DSL} - \mu_{NDSL}$

귀무가설, $H_0 : \mu_{DSL} = \mu_{NDSL},\ \mu_{DSL} - \mu_{NDSL} = \mu_D = 0$
대립가설, $H_1 : \mu_{DSL} > \mu_{NDSL},\ \mu_{DSL} - \mu_{NDSL} = \mu_D > 0$

3-1 (b) : 유의수준 5%에서의 가설검정을 수행하고 결론을 도출하시오.

기초 통계량 :

n = 101, 유의수준 = 0.05

$\overline{DSL} = 178.5,\ \sigma_{DSL} = 7.05$

$\overline{NDSL} = 179.9,\ \sigma_{NDSL} = 7.05$

$\overline{D} = \overline{DSL} - \overline{NDSL} = 178.5 - 179.9 = -1.4$

${\sigma_D}^2 = Var(\overline{D}) = Var(\overline{DSL} - \overline{NDSL}) = Var(\overline{DSL}) + Var(\overline{NDSL}) - 2Cov(\overline{DSL}, \overline{NDSL})$

$= Var(\overline{DSL}) + Var(\overline{NDSL}) = 7.05^2 + 7.05^2 = 99.405$

$\sigma_D = \sqrt{99.405} = 7.05\sqrt{2}$

데이터 수집 때 DSL 사람들과 아닌 사람들을 따로 구했기 때문에 독립이라고 바라봐도 된다.

검정 통계량 :

모평균에 대한 가설을 세웠기 때문에 t-분포 혹은 정규분포를 세워야 하지만 n 의 값이 101 으로 30 보다 훨씬 큰 값이기 때문에 중심극한정리에 의해서 정규분포를 검정통계량으로 세워도 된다. 검정통계량은 Z^* 로 다음과 같다.

$$Z^* = \frac{\overline{D} - \mu_D}{\sigma_D/\sqrt{n}} = \frac{-1.4 - 0}{7.05\sqrt{2}/\sqrt{101}} = \frac{-1.4\sqrt{101}}{7.05\sqrt{2}} \sim N(0,1)$$

기각역 :

$$Z^* > Z_\alpha = Z_{0.05} = 1.6449$$

위의 식이 성립된다면 우리는 대립가설을 기각하게 된다.

$$Z^* = \frac{-1.4\sqrt{101}}{7.05\sqrt{2}} = -1.4112$$

1.6449 보다 훨씬 작기 때문에 식이 성립하지 않게 되며 대립가설을 기각하지 못하게 된다.

통계적 결론 :

유의수준 0.05 에 의해서 DSL 학회원들의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 크다고 할 수가 없다.