기수 : 10기　　이름 : 연제명

문제 1　Central Limit Theorem

중심극한정리는 확률변수의 합 형태(sum of random variables)의 극한분포를 손쉽게 구할 수 있도록 해주기에 통계학에서 가장 자주 사용하는 정리입니다. 이 문제에서는 중심극한정리의 정의와 그 활용에 대해 짚어보겠습니다.

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

> ☝ 통계학입문(3판) 7장 참고
> ☝ Hogg(8판) 4장 2절, 5장 3절 참고

(Answer)

- **정의**

독립적인 확률 변수들의 합 또는 평균은 샘플의 크기가 충분히 크다면, 정규 분포에 근사하게 되는 현상을 설명하는 수학적 정리

- **의미**

(1) 표본 크기가 증가할수록 표본 평균이 모평균에 수렴한다는 점이 있다. 이는 통계 추론에서 신뢰성 있는 결과를 얻을 수 있도록 한다.

(2) 많은 통계적 추론 방법의 기초로 사용된다. 예를 들어, 가설 검정, 신뢰 구간 추정, 회귀 분석 등에 이용하여 모집단에 대한 정보를 얻을 수 있다.

(3) 정규 분포를 가정하는 많은 통계 모델들을 합리적으로 사용할 수 있게 한다.

(4) 비교적 적은 가정과 표본 크기로부터 모집단에 대한 통계적 추론을 수행할 수 있게 한다.

(1-2) ✎ 중심극한정리가 통계적 추론 중 "구간추정"에서 어떻게 유용한지 서술하시오.

> ☝ Hogg(8판) 4장 2절 참고

(Answer)

- **구간추정**

모집단의 파라미터(ex. 평균, 분산)에 대한 추정치를 구간으로 제공함으로써, 추정치의 신뢰성과 불확실성을 동시에 고려하는 방법

- **구간추정에서 중심극한정리의 유용성**

(1) 정규분포 가정

중심극한정리를 기반으로 구간추정을 수행할 때, 표본 평균 또는 표본 비율과 같은 추정치가 정규 분포를 따른다고 가정할 수 있다. 이를 통해 정규 분포에 관한 추론 방법을 적용할 수 있다.

(2) 신뢰 구간 계산

중심극한정리에 따라 표본 평균 또는 표본 비율의 분포는 표본 크기가 충분히 크다면, 정규 분포에 가까워진다. 이를 이용하여 구간추정을 수행할 때, 표본 평균 주변의 신뢰 구간을 계산할 수 있다. 예를 들어, 표본 평균을 중심으로 표준 오차와 신뢰 수준에 따라 상한과 하한을 계산하여 신뢰 구간을 구성할 수 있다.

(3) 신뢰성 평가

중심극한정리를 이용한 구간추정은 표본 크기가 충분히 크다면, 추정치 주변의 신뢰 구간을 모수의 실제 값이 포함된 확률을 나타내는 신뢰 수준으로 해석할 수 있다. 예를 들어, 95% 신뢰 수준의 구간추정은 표본을 반복해서 추출한다면, 이러한 구간들이 95%의 경우 실제 모수를 포함한다는 의미다.

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

☝ 표본분산 s^2는 모분산 σ^2로 확률수렴한다는 사실을 이용할 수 있습니다. ☝ Slutsky's theorem을 이용할 수 있습니다. ☝ Hogg(8판) 5장 1~3절 참고

(Answer)

• 중심극한정리에 따르면, 모집단이 어떤 분포를 가지더라도, 표본의 크기가 충분히 크다면 표본평균은 정규분포를 따른다고 할 수 있다. 여기서, X1, X2, ..., Xn은 독립적으로 동일한 확률변수를 표본으로 추출한 값들이고, μ는 모평균, σ^2는 모분산이다.

(X1 + X2 + ... + Xn) / n ~ N(μ, σ^2/n)

• 표본분산 s^2은 표본의 크기가 커질수록 모분산 σ^2에 확률적으로 수렴한다. 여기서 ε는 어떤 양수이다.

lim(n->∞) P(|s^2 - σ^2| < ε) = 1

• 모분산 σ^2에 대한 표준오차는 sqrt(σ^2/n)으로 나타낼 수 있다. 표본분산 s^2 역시 표본의 크기가 충분히 크다면 모분산 σ^2에 확률적으로 수렴하므로, 표준오차와 표본분산은 확률적으로 동일한 분포를 가진다.

SE = σ/√n

• Slutsky의 정리에 따르면, 확률로 수렴하는 두 가지 변수 중 하나가 상수이고 다른 하나가 확률변수일 때, 이 두 가지의 조합은 확률적으로 동일한 분포를 따른다.

표준오차(SE)는 모분산(σ^2)을 표본의 크기(n)로 나눈 값으로 나타낼 수 있다. 표본분산(s^2)은 모분산(σ^2)에 확률적으로 수렴하므로, Slutsky의 정리에 따라 다음과 같이 표현할 수 있다:

SE ≈ s/√n

∴ 따라서 표준오차(SE) 부분의 모분산(σ^2)을 표본분산(s^2)으로 대체할 수 있다.

문제 2	Student's Theorem

스튜던트 정리는 통계적 추정에서 필요한 정리 중 하나로, 표본평균과 표본분산이 어떤 분포를 갖는지 알려줍니다. 이 문제에서는 스튜던트 정리의 내용을 어떻게 수식적으로 유도할 수 있는지 짚어보겠습니다.

스튜던트 정리는 다음과 같이 총 4개의 내용으로 구성되어 있습니다.

① $\overline{X} \approx N(\mu, \frac{\sigma^2}{n})$ → 〈문제 1〉에서 증명함

② 표본평균 $\overline{X}$ 와 표본분산 s^2은 서로 독립이다.

③ ???

④ ???

(2-1) ✎ ③의 내용을 쓰고 증명하시오.

☝ Hogg(8판) 3장 6절 참고

☝ 무작위표본 $X_1, \cdots, X_n$이 독립적으로 동일하게(independently and identically distributed) 평균이 μ이고 분산이 σ^2인 정규분포를 따를 때, 자유도가 n인 카이제곱분포를 따르는 새로운 확률변수 V를 아래와 같이 두어 증명에 활용할 수 있습니다.

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

(Answer)

• 먼저 주어진 정보에 따라 표본의 합을 표준화한다. 표준화는 각 표본에서 평균을 빼고, 표준편차로 나누는 과정을 말한다.

Z = (X - μ) / σ

• 표본의 합을 표준화하여 얻은 새로운 확률변수 Z를 사용하여 V를 다음과 같이 정의한다.

V = Σ (Zi)^2

• Z를 정의한 것과 같은 방법으로, 각 표본을 표준화하고 표준화된 표본을 제곱한 값으로 이루어진 V는 자유도가 n인 카이제곱분포를 따른다.

V ~ x^2(n)

(2-2) ✎ ④의 내용을 쓰고, (2-1)을 이용하여 증명하시오.

☝ Hogg(8판) 3장 6절 참고

☝ t분포의 정의에 따르면, 표준정규분포를 따르는 확률변수와 카이제곱분포를 따르는 확률변수를 이용하여 t분포를 유도할 수 있습니다.

(Answer)

• 먼저 Z와 V를 독립적으로 가정한다. 표준정규분포를 따르는 확률변수 Z를 통해 새로운 확률변수 T를 다음과 같이 정의한다.

T = Z / sqrt(V/n)

• 각 확률변수 Z와 V를 독립적으로 가정하고 T를 정의하였으므로, T는 자유도가 n인 t분포를 따른다.

문제 3	t-test

t검정은 모집단이 정규분포를 따르지만 모표준편차를 모를 때, 모평균에 대한 가설검정 방법입니다. 대개 두 집단의 모평균이 서로 차이가 있는지 파악하고자 할 때 사용하며, 표본평균의 차이와 표준편차의 비율을 확인하여 통계적 결론을 도출합니다.

(3-1) 어떤 학우가 DSL 학회원(동문 포함)의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 크다는 주장을 하여, 실제로 그러한지 통계적 검정을 수행하려고 합니다. 며칠간 표본을 수집한 결과 다음의 결과를 얻었다고 합시다.

표본 수 : 각 101명
측정에 응한 DSL 학회원들의 평균 키 : 178.5cm / 표준편차 : 7.05cm
측정에 응한, DSL 학회원이 아닌 사람들의 평균 키 : 179.9cm / 표준편차 : 7.05cm

(a) ✎ 귀무가설과 대립가설을 설정하시오.

(Answer)

- **귀무가설(H_0):** DSL 학회원의 평균 키는 DSL 학회원이 아닌 사람의 평균 키와 같다.
- **대립가설(H_1):** DSL 학회원의 평균 키는 DSL 학회원이 아닌 사람의 평균 키보다 크다.

(b) ✎ 유의수준 5%에서의 가설검정을 수행하고 결론을 도출하시오.

☝ 통계학입문(3판) 7장 참고
☝ 어떤 검정통계량이 어떤 분포를 따르는지, 언제 귀무가설을 기각하는지 정해야 합니다.

(Answer)

- DSL 학회원들의 표본 평균 ($\bar{x}_1$) = 178.5 cm

  DSL 학회원들의 표본 표준편차 (s_1) = 7.05 cm
- DSL 학회원이 아닌 사람들의 표본 평균 ($\bar{x}_2$) = 179.9 cm

  DSL 학회원이 아닌 사람들의 표본 표준편차 (s_2) = 7.05 cm
- 표본 크기 ($n_1 = n_2$) = 101
- 표본 편차의 합산 값 $(s_1^2/n_1) + (s_2^2/n_2) = (7.05^2/101) + (7.05^2/101) = 0.495$
- 검정통계량 $(t) = (\bar{x}_1 - \bar{x}_2) / \text{sqrt}((s_1^2/n_1) + (s_2^2/n_2))$

  $= (178.5 - 179.9) / \text{sqrt}(0.495) = -1.4 / 0.703 = -1.99$
- 자유도(df) = $n_1 + n_2 - 2 = 101 + 101 - 2 = 200$
- 자유도 200에 대한 양측 검정에서 유의수준 5%에 해당하는 임계값은 약 ±1.972
- $-1.99 < -1.972$

∴ 검정통계량 (t)는 임계값을 넘지 못하므로, 귀무가설을 기각할 수 없다. 따라서 "DSL 학회원의 평균 키는 DSL 학회원이 아닌 사람의 평균 키보다 크다"는 통계적으로 유의하지 않다.