23 – 2 DSL 정규 세션

기초과제 1 통계적 사고

2019136008 김민석

1-1. 중심극한정리(Central Limit Theorem)의 정의를 서술하시오.

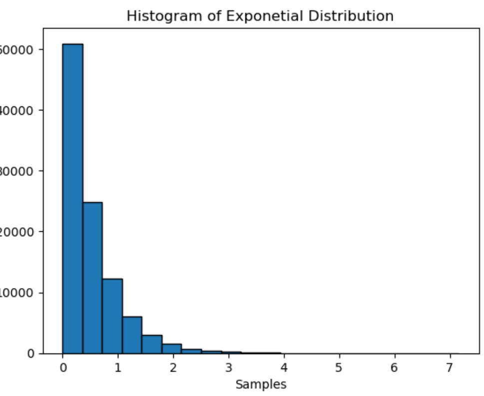
- X_i 는 서로 독립적이고 동일한 분포를 갖고 (i.i.d.), $Var(X_i)$ 가 유한한 값을 갖는 확률 변수 X 에 대해 N개의 표본의 합의 표준화 된 값, 즉 $\frac{(X_1 + X_2 + X_3 + \cdots X_N) - \mu \cdot N}{\sigma \cdot \sqrt{N}}$ 이 두 상수 a, b 사이에 있을 확률은 N을 무한으로 보냈을 때 표준 정규분포 함수 $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ 의 구간 (a, b) 의 적분과 수렴한다.

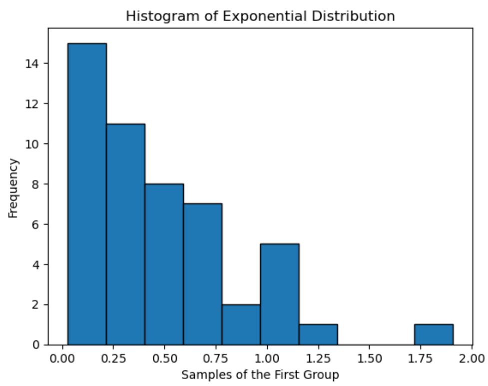
$$\therefore \text{For i.i.d. } X_i,\ 0 < Var(X_i) < \infty : \lim_{N \to \infty} P\left(a < \frac{(X_1 + X_2 + X_3 + \cdots X_N) - \mu \cdot N}{\sigma \cdot \sqrt{N}} < b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

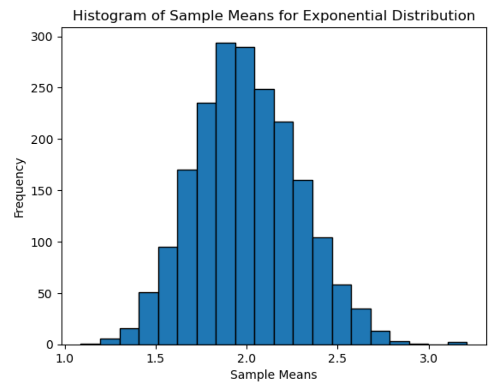
1-2. : 중심극한정리가 통계적 추론 중 “구간추정”에서 어떻게 유용한지 서술하시오.

표본평균의 분포를 정규 분포로 근사하여 표본 평균에 대한 신뢰구간을 구할 수 있다. 신뢰구간 내에 모집단의 평균이 어느정도의 정확도로 존재하는지 추정할 수 있으므로 구간 추정이 유용하게 쓰인다.

1-3 © : .ipynb 파일에서 $Unif$ (0, 1) 의 분포에 대해서 중심극한정리가 적용되는 예시가 있습니다. 코드를 참조하면서 지수분포인 Exp(2) 분포에 대해서 중심극한정리가 적용되는 모습을 보이시오.







2.1 다음과 같은 정방행렬에 Diagonalization 을 적용시켜서 나오게 되는 대각행렬을 쓰시오.

a) $\begin{bmatrix} 6 & -1 \\ 2 & 3 \end{bmatrix}$

$D = \begin{pmatrix} 4 & 0 \\ 0 & 5 \end{pmatrix}$ $\quad P = \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}$

$(6-\lambda)(3-\lambda) - (-1)(2) = 0$

$\lambda^2 - 9\lambda + 18 + 2$

$\lambda^2 - 9\lambda + 20$ $\quad \lambda = \overset{i)}{4}$ or $\overset{ii)}{5}$

i) $\left(\begin{array}{cc|c} 2 & -1 & 0 \\ 2 & -1 & 0 \end{array}\right)$ $\quad$ ii) $\left(\begin{array}{cc|c} 1 & -1 & 0 \\ 2 & -2 & 0 \end{array}\right)$

$\left(\begin{array}{cc|c} 2 & -1 & 0 \\ 0 & 0 & 0 \end{array}\right)$ $\quad \left(\begin{array}{cc|c} 1 & -1 & 0 \\ 0 & 0 & 0 \end{array}\right)$

$2x - y = 0$ $\begin{pmatrix} t \\ 2t \end{pmatrix}$ $\quad x - y = 0$ $\begin{pmatrix} t \\ t \end{pmatrix}$

b) $\begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 1 \\ -1 & 0 & 1 \end{bmatrix}$

$\det \begin{pmatrix} 2-\lambda & 0 & 0 \\ 1 & 2-\lambda & 1 \\ -1 & 0 & 1-\lambda \end{pmatrix} = (2-\lambda)(2-\lambda)(1-\lambda) = 0$

$\lambda = 2$ or 1

$\lambda=2)$ $\begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ -1 & 0 & -1 \end{pmatrix}$ $\quad \lambda=1)$ $\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ -1 & 0 & 0 \end{pmatrix}$

$\begin{bmatrix} k \\ t \\ -k \end{bmatrix}$ $\quad \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$

$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$ $\quad \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$

$D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ $\quad P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & -1 & -1 \end{bmatrix}$

c) $\begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ $\quad (\lambda-3)(\lambda-1)(\lambda-2) = 0$

$\lambda=3)$ $\begin{bmatrix} 0 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ $\quad \lambda=1)$ $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ $\quad \lambda=2)$ $\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

$\begin{bmatrix} t \\ 0 \\ 0 \end{bmatrix}$ $\quad \begin{bmatrix} 0 \\ t \\ 0 \end{bmatrix}$ $\quad \begin{bmatrix} 0 \\ 0 \\ t \end{bmatrix}$

$D = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ $\quad P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

2.2 다음과 같은 행렬에 SVD 를 적용시켜서 나오게 되는 ∑ 행렬을 구하시오.

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad AA^T = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\det \begin{bmatrix} 1-\lambda & 1 & 0 \\ 1 & 1-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{bmatrix} \qquad (1-\lambda)(1-\lambda)^2 - (1-\lambda) = 0$$

$$(1-\lambda)(\lambda-2)\lambda = 0 \qquad \overset{i)}{\lambda = 2} \text{ or } \overset{ii)}{\lambda = 1} \text{ or } \overset{iii)}{\lambda = 0}$$

$$\det\begin{pmatrix} 2-\lambda & 0 \\ 0 & 1-\lambda \end{pmatrix} = (2-\lambda)(1-\lambda) = 0 \qquad \overset{i)}{\lambda = 2} \text{ or } \overset{ii)}{1}$$

i) $\begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$ $\begin{bmatrix} k \\ k \\ 0 \end{bmatrix}$ ii) $\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ $\begin{bmatrix} 0 \\ 0 \\ t \end{bmatrix}$ iii) $\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ $\begin{bmatrix} k \\ -k \\ 0 \end{bmatrix}$

i) $\begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$ $\begin{bmatrix} t \\ 0 \end{bmatrix}$ ii) $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ $\begin{bmatrix} 0 \\ t \end{bmatrix}$

$$U_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \quad U_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad U_3 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix} \qquad U_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad U_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

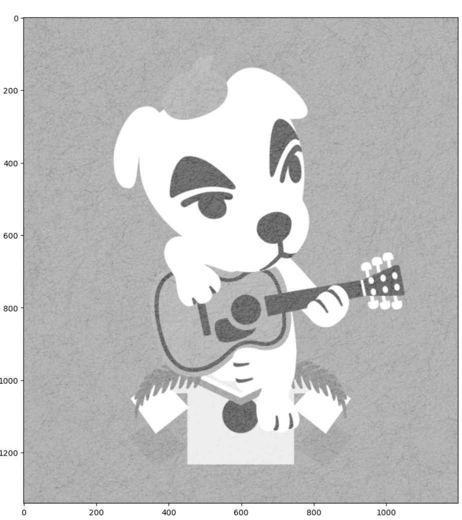
$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

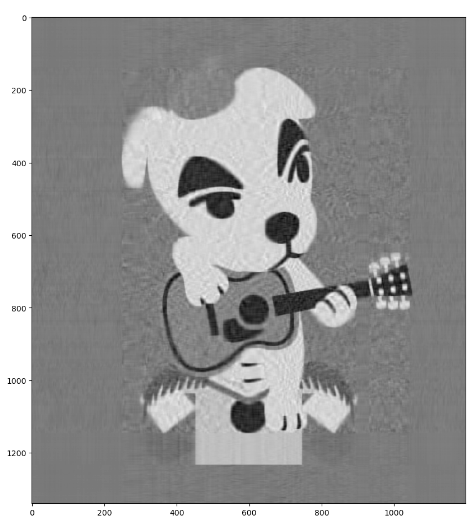
$$V = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix}$$

2-3 © : .ipynb 파일에서는 SVD (Singular Value Decomposition, 특이값-분해) 를 실제로 이미지 압축을 위해서 활용하는 예시를 보여주고 있습니다. 해당 코드를 본 뒤에 새로운 사진에 대해서 원본에 비해서 적은 용량을 차지하면서도 원본에 대한 정보를 유지해주는 차원 수가 무엇인지 알아냅시오





```
Image Float uses  3232002  Non-zero Elements
Image Composition uses  177730  Non-zero Elements
Processing Time for 10 Full Images: 9.29379 seconds
Processing Time for 10 Compressed Images: 0.36467 seconds
```

스튜던트 정리는 다음과 같이 총 4 개의 내용으로 구성되어 있습니다.

① $\overline{X} \approx N(\mu, \sigma^2/n)$

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

③ $(n-1)S^2/\sigma^2$ 는 $\mathcal{X}^2(n-1)$ 분포를 따른다.

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

3-1 :③ 에 있는 내용을 증명하시오.

$$
\begin{aligned}
V &= \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \\
&= \sum \left(\frac{(X_i - \bar{X}) + (\bar{X} - \mu)}{\sigma} \right)^2 \\
&= \sum \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + n \cdot \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 + \sum 2 \frac{(X_i - \bar{X})(\bar{X} - \mu)}{\sigma^2} \\
&= \sum \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \quad \left(\because \ \sum (X_i - \bar{X}) = (n-1)(\bar{X} - \bar{X}) = 0 \right) \\
&= \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2
\end{aligned}
$$

②에의해 S^2과 $\bar{X}$는 독립이므로 양변에 적률생성함수 mgf 를 취해주면

$$(1-2t)^{-n/2} = E\left(\exp\left(t(n-1)S^2/\sigma^2\right)\right) \cdot (1-2t)^{-1/2}$$

따라서 $(n-1)S^2/\sigma^2$ 의 적률생성함수는 자유도 $(n-1)$의 카이제곱 분포 임을 알 수 있다.

$$\therefore \ (n-1)S^2/\sigma^2 \approx \chi^2(n-1)$$

3-2 : ④에 있는 내용을 증명하시오.

.

$$
\begin{aligned}
T &= \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \\
&= \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}(n-1)\frac{S^2}{\sigma^2}}}
\end{aligned}
$$

Consider $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$ ($\because$ ③)

We can rewrite $T = \frac{W}{\sqrt{V/(n-1)}}$

rmk $t(r) = \frac{W}{\sqrt{(V/r)}}$

$T \sim t(n-1)$

4-1 : 어떤 학우가 DSL 학회원(동문 포함)의 평균 키가 DSL 학회원이 아닌 사람의 평균 키보다 크 다는 주장을 하여, 실제로 그러한지 통계적 검정을 수행하려고 합니다. 며칠간 표본을 수집한 결과 다음의 결과를 얻었다고 합시다.

표본 수 : 총 210 명, 각 105 명

측정에 응한 DSL 학회원들의 평균 키 : 173.5cm / 표준편차 : 7.05cm

측정에 응한, DSL 학회원이 아닌 사람들의 평균 키 : 171.4cm / 표준편차 : 7.05cm

a) 귀무가설과 대립가설을 설정하시오.

$H_0 : \mu_1 - \mu_2 \le 0 \quad H_1 : \mu_1 - \mu_2 > 0$

b) 유의수준 5%에서의 가설검정을 수행하고 결론을 도출하시오.

i) 등분산 $\sigma_1 = \sigma_2$ 일경우 가정

$\mu_1 - \mu_2$ 의 검정통계량 $T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \ge \dfrac{2.1}{7.05 \cdot \sqrt{\frac{2}{105}}}$ (cond'T where $\mu_1 = \mu_2$)

이때 각 $n_1 = n_2 = 105 > 30$ 이므로 By CLT

$\sim N(0,1)$

$z_{0.05} = 1.645 \quad T \ge 2.15829 > 1.64$

$\rightarrow$ reject H_0 accept H_1

Similarly 등분산이 아닐 때에도 $n_1 = n_2$ 이기때문에 같은 결과가 나온다.

4-2 © : 또 다른 학우가 다른 학회인 ESC 의 키들도 포함이 된다고 알려주었으며, 새로운 데이터를 heights.csv 파일에 저장해놓았다고 합니다. 이 학우는 학회마다의 평균 키가 똑같다는 주장을 하고 있으며, 해당 학우가 준 데이터를 통해서 이 주장을 검정하려고 합니다.

a) 귀무가설과 대립가설을 설정하시오.

$H_0 : \mu_1 = \mu_1 = \mu_2 \qquad H_1 : \mu_1 \ne \mu_2$ or $\mu_1 \ne \mu_3$ or $\mu_2 \ne \mu_3$

b) © 파이썬의 scipy.stats 을 활용해서 유의수준 5%에서의 가설검정을 수행하고 결론을 도출하시오. 결론은 .ipynb 파일에 쓰셔도 괜찮습니다.

```
f_statistic, p_value = stats.f_oneway(DSL.iloc[:, 0], ESC.iloc[:, 0], Else.iloc[:, 0])

print(f'F-statistic: {f_statistic}')
print(f'P-value: {p_value}')
✓ 0.0s

F-statistic: 6.82185346949053
P-value: 0.0013059664602332191
```

```
alpha = 0.05

if p_value < alpha:
    print("유의수준 5%에서 귀무가설을 기각합니다.")
else:
    print("유의수준 5%에서 귀무가설을 기각하지 않습니다.")
✓ 0.0s

유의수준 5%에서 귀무가설을 기각합니다.
```