인공지능 수학: 통계학
5.표본분포

윤상민

E-mail : smyoon@kookmin.ac.kr
Office : 02-910-4645

서론

- 통계적 추론
  - 표본 조사를 통해 모집단에 대한 해석을 진행
  - 전수조사는 실질적으로 불가능한 경우가 많음
- 표본 조사는 반드시 오차가 발생
  - 적절한 표본 추출 방법 필요
  - 표본과 모집단과의 관계를 이해해야 함

서론

- 단순랜덤추출법 (random sampling)
- 난수표 사용
- 랜덤넘버 생성기 사용
  - https://colab.research.google.com/

  - import random
  - [random.randint(1, 10) for i in range(10)]

표본분포

- 표본조사를 통해 파악하고자 하는 정보:
  - 모수 (Parameter)
- 모수의 종류:
  - 모평균, 모분산, 모비율 등
  - 모수를 추정하기 위해 표본을 선택하여 표본 평균이나 표본 분산등 계산
- 통계량 (statistic):
  - 표본 평균이나 표본 분산과 같은 표본의 특성값

표본분포

- 50만명의 전국 고등학교 1학년 학생의 키를 조사하기 위해 1000명을 표본 조사함
  - 표본의 평균 계산
  - 표본의 평균은 표본의 선택에 따라 달라짐
  - 따라서 표본평균은 확률변수
- 표본 평균이 가질 수 있는 값도 하나의 확률분포를 가짐
  - 그 분포가 무엇인지가 표본을 해석하는데 있어서 매우 중요함
- 통계량의 확률분포: 표본분포 (sampling distribution)

표본분포

- 표본평균
  - 모평균을 알아내는데 쓰이는 통계량
- 표본평균의 분포
  - $x_1, x_2, \ldots, x_n$
    - 평균: μ, 분산: σ^2
    - **정규모집단에서** 추출된 표본의 측정값
  - 표본평균
    - $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
    - $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

표본분포

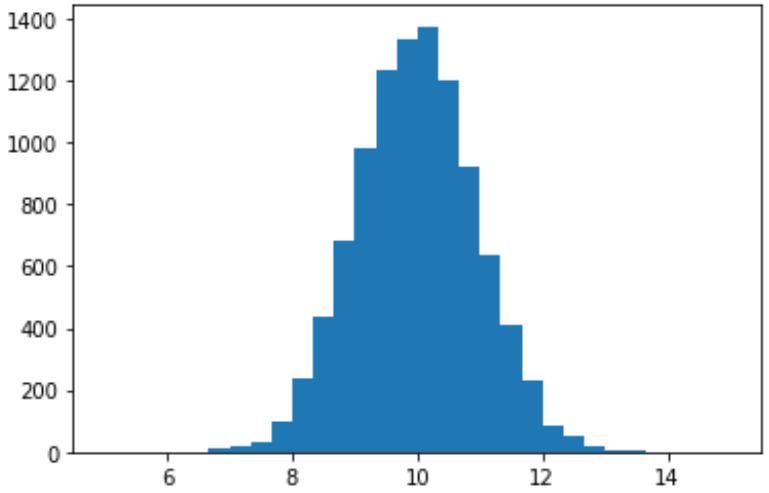
- import numpy as np
- xbars = [np.mean(np.random.normal(size = 10)) for i in range(10000)]
- print("mean %f, var %f" %(np.mean(xbars), np.var(xbars)))

$$\mu = 0, \sigma = 1$$
$$n = 10$$
$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{1}{10}$$

표본분포

- xbars = [np.mean(np.random.normal(loc=10, scale=3, size = 10)) for i in range(10000)]
- print("mean %f, var %f" %(np.mean(xbars), np.var(xbars)))
- import matplotlib as plt
- h=plt.pyplot.hist(xbars, range=(5,15), bins=30)

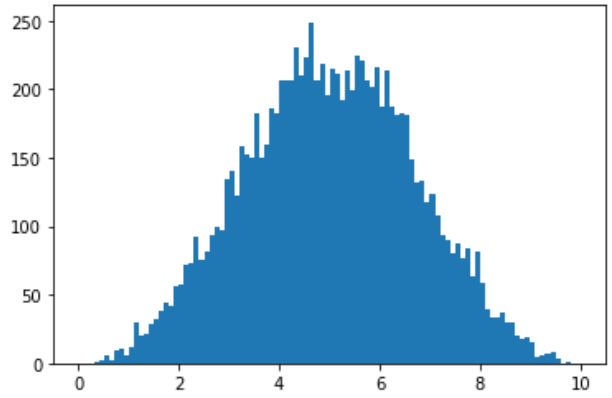
$$\mu = 10$$
$$\sigma = 3$$
$$n = 10$$



중심극한정리

- 중심극한정리(central limit theorem)
  - $x_1, x_2, \ldots, x_n$
    - 평균: μ, 분산: σ^2
    - **~~정규~~모집단에서** 추출된 표본의 측정값
  - 표본평균
    - $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
    - n이 충분히 큰 경우 $(n \geq 30)$,
      - **근사적**으로 $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

중심극한정리



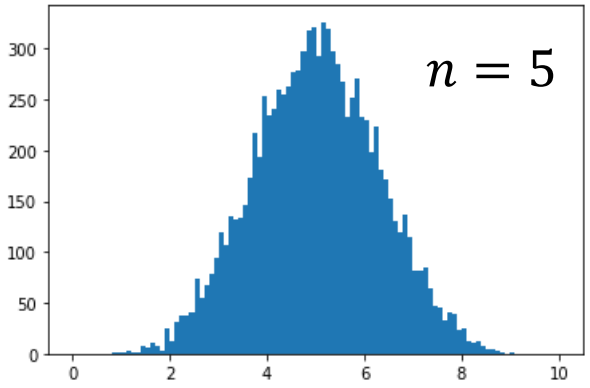
- import numpy as np
- import matplotlib as plt
- n=3
- xbars=[np.mean(np.random.rand(n) * 10) for i in r ange(10000)]
- print("mean %f, var %f" %(np.mean(xbars), np.var(xbars)))
- h=plt.pyplot.hist(xbars, range=(0,10), bins=100)

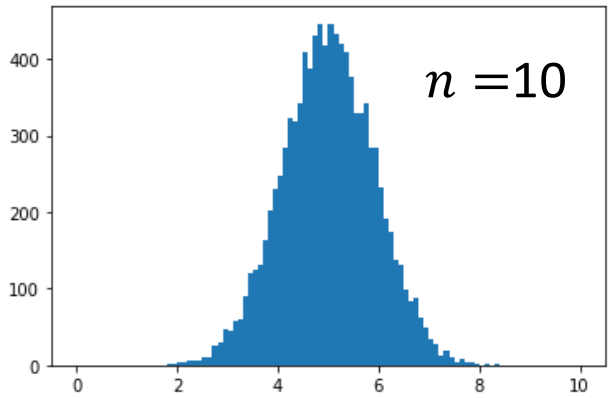
Uniform Distribution

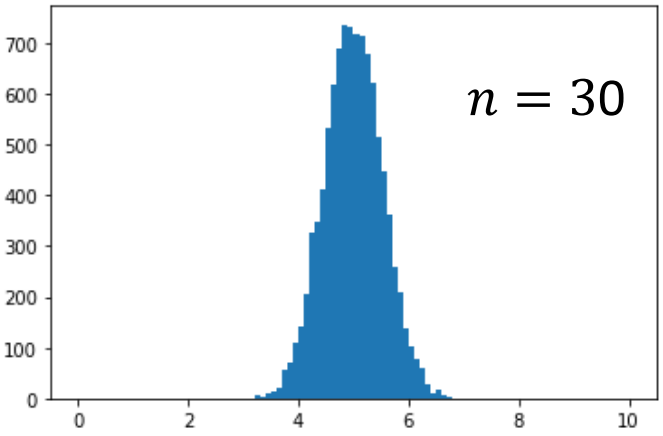
$$E(X) = 5$$

$$\mathrm{Var}(X) = \frac{(10-0)^2}{12} = 8.3$$

중심극한정리


$n = 5$
300
250
200
150
100
50
0
0
2
4
6
8
10


n =10
400
300
200
100
0
0
2
4
6
8
10


$n = 30$
700
600
500
400
300
200
100
0
0
2
4
6
8
10