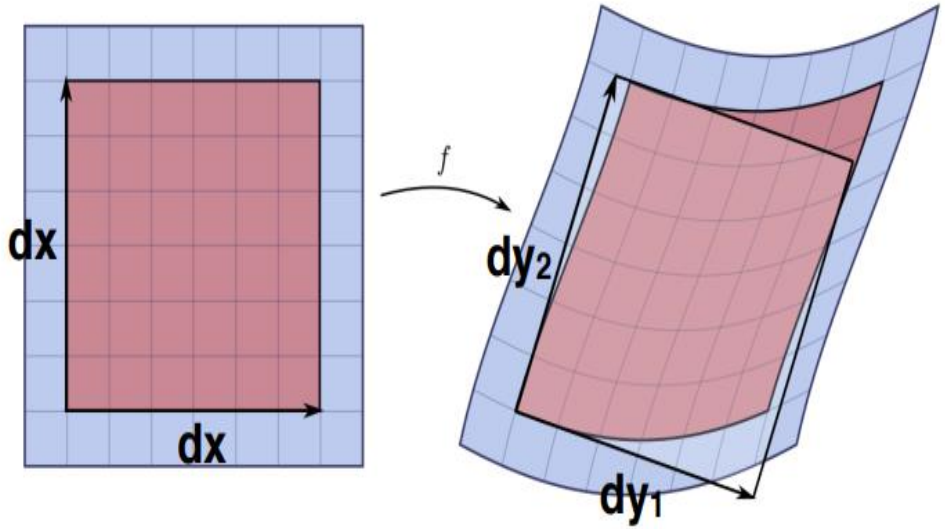
Effective Random Generator and Central Limit Theorem

2017135002/최성윤

기본원리(2차원 확률변환)

이전에 Return Method를 사용한 난수 발생기는 비효율적이다. (사용하지 않는 난수가 많다.)

-> Jacobian을 활용한 2차원 확률변환



1 과 2를 연립 한다면

$$\frac{p(y_1, y_2)}{p(x_1, x_2)} = \quad p(y_1, y_2) = \left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right|$$

1.확률보전법칙

2차원에서도 확률보존법칙은 성립이 된다.

$$p(x_1, x_2)dx_1dx_2 = p(y_1, y_2)dy_1dy_2$$

$p(x_1, x_2) = 1$을 만족한다.(균일분포)

$$\mathrm{d}x_1\,\mathrm{d}x_2 = \mathrm{p}(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2$$

2.Jacobian

2차원에서도 확률보존법칙은 성립이 된다.

$$dx_1dx_2 = dy_1dy_2 \left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right| \qquad \left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right| = \frac{\partial x_1}{\partial y_1}\frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2}\frac{\partial x_2}{\partial y_1}$$

목적확률분포는 Jacobian과 동일하다

기본원리(2차원 확률변환)

발생시킨 난수 x_1, x_2 (0~1)을 Gaussian 분포로 변환을 한다.

$$y_1 = \sqrt{-2\ln x_1}\cos 2\pi x_2$$
$$y_2 = \sqrt{-2\ln x_1}\sin 2\pi x_2$$

Jacobian에 대입

$$\left|\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}\right| = \frac{\partial x_1}{\partial y_1}\frac{\partial x_2}{\partial y_2} - \frac{\partial x_1}{\partial y_2}\frac{\partial x_2}{\partial y_1}$$

$$\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = + \left[\frac{1}{\sqrt{2\pi}}e^{-y_1^2/2}\right]\left[\frac{1}{\sqrt{2\pi}}e^{-y_2^2/2}\right]$$

y_1, y_2가 독립적으로 Gaussian 분포를 가지게 된다

$$\left[\frac{1}{\sqrt{2\pi}}e^{-y_1^2/2}\right]\left[\frac{1}{\sqrt{2\pi}}e^{-y_2^2/2}\right]dy_1dy_2 = p(y_1,y_2)dy_1dy_2 = p(y_1)p(y_2)dy_1dy_2$$

$\mathrm{p}(y_1), p(y_2)$ 각각 Gaussian 분포

실행

```
def mygauss(m,std,N):
    #난수 N/2개씩 생성하여 총 N개 생성
    # float object cannot be interpreted as an integer 오류 발생
    x1=np.array(myran(int(N/2)))
    x2=np.array(myran(int(N/2)))
    #변환 (표준정규분포의 형태를 가진다. 평균 0 , 표준편차 1)
    y1=np.sqrt(-2*np.log(x1))*np.cos(2*np.pi*x2)
    y2=np.sqrt(-2*np.log(x1))*np.sin(2*np.pi*x2)
    #평균 m, 표준편차 std를 가지는 정규분포 형성
    Y1=std*y1+m
    Y2=std*y2+m
    Y3=[]
    #Y1,Y2를 합친 리스트 만들기
    for i in range(5000):
        Y3.append(Y1[i])
        Y3.append(Y2[i])
    #Y1,Y2,Y3히스토그램 비교하깅
    plt.hist(Y1, bins=100,histtype='step',label='Y1')
    plt.hist(Y2, bins=100,histtype='step',label='Y2')
    plt.hist(Y3, bins=100,label='Y1+Y2',density='True')
    plt.xlabel("Random Number")
    plt.ylabel("Number")
    plt.legend()
    return Y3
```

총 만들고자 하는 난수 개수 N개에서 각각 N/2개씩 생성

$$y_1 = \sqrt{-2\ln x_1}\cos 2\pi x_2$$

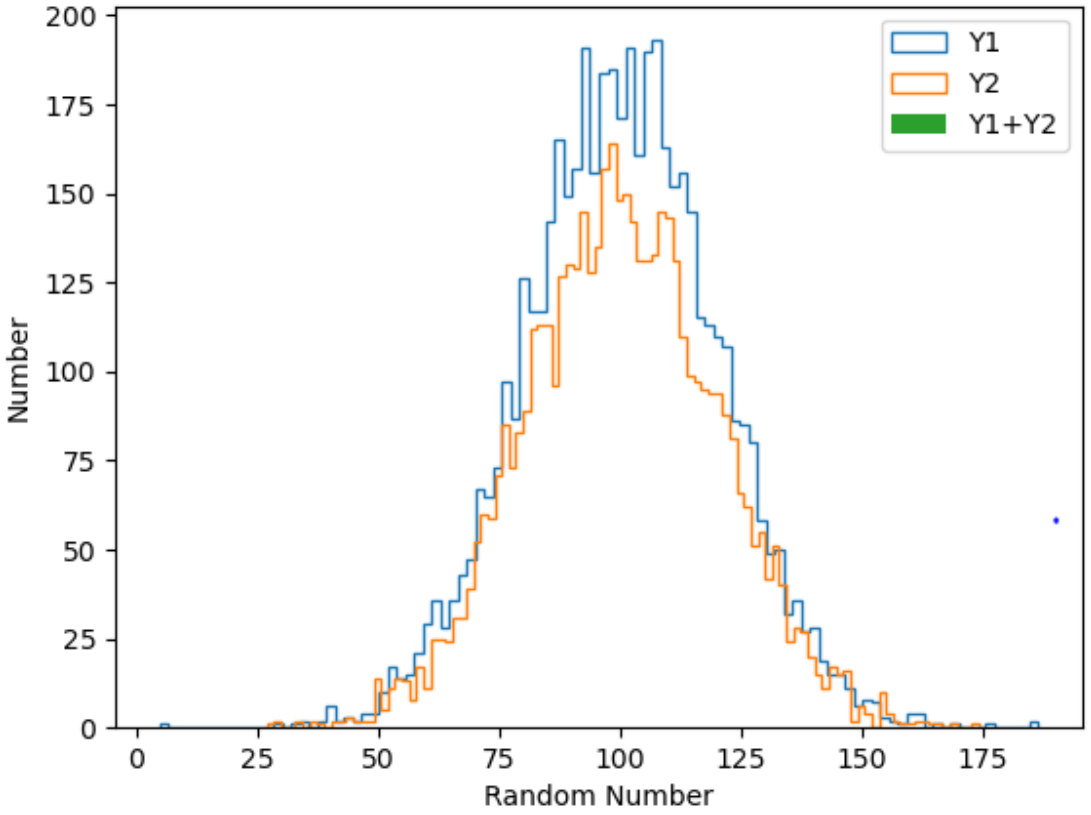
$$y_2 = \sqrt{-2\ln x_1}\sin 2\pi x_2$$

다음 식을 적용하여 y_1, y_2 변환

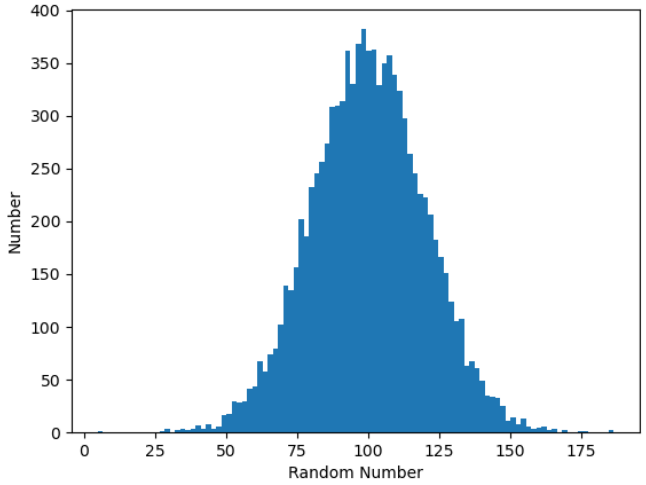
평균 m 표준편차 20을 가지는 정규분포로 변환

y1,y2 배열을 합하여 Y3 생성
N개의 난수 생성

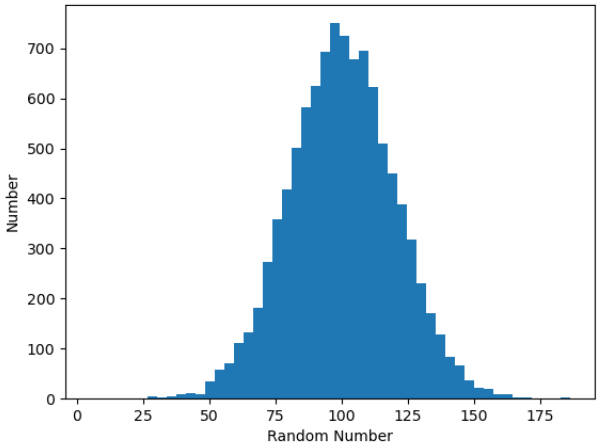
결과



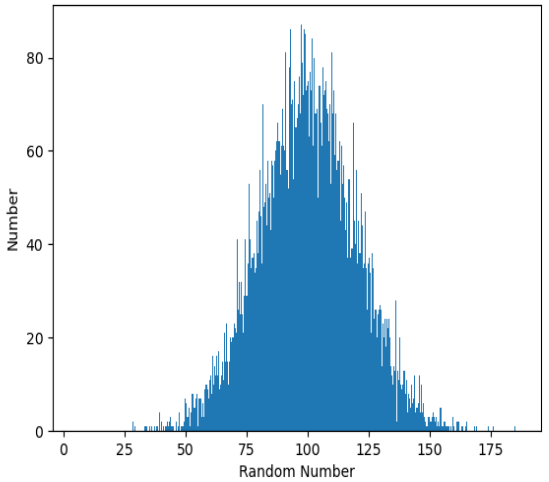
Bins=100



Bins=50



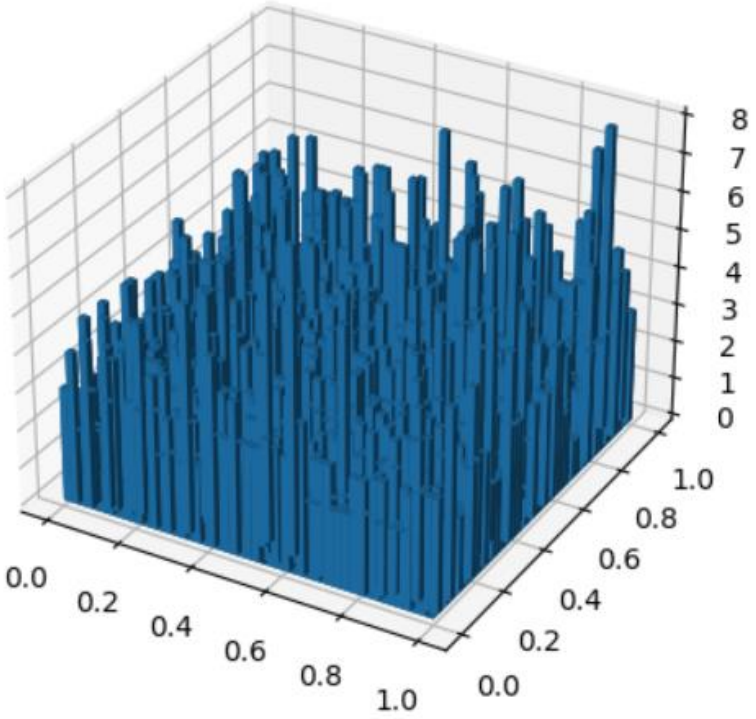
Bins=500



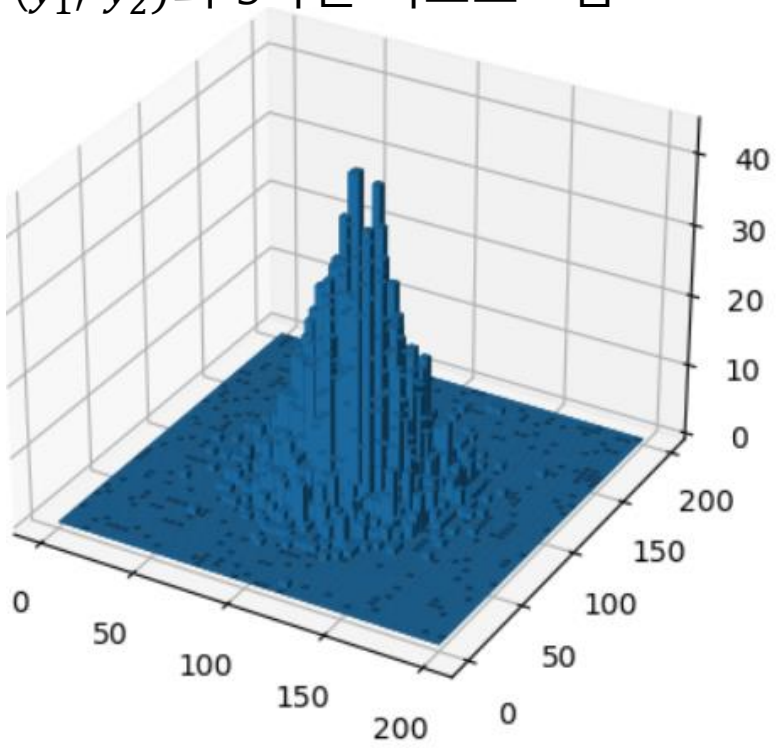
토의사항

1. (x_1, x_2) (y_1, y_2) 분포 비교

(x_1, x_2)의 3차원 히스토그램

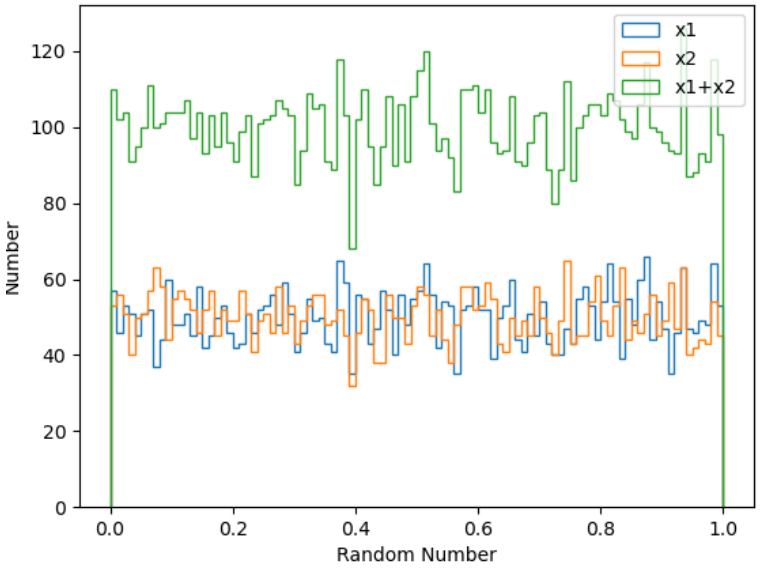


(y_1, y_2)의 3차원 히스토그램

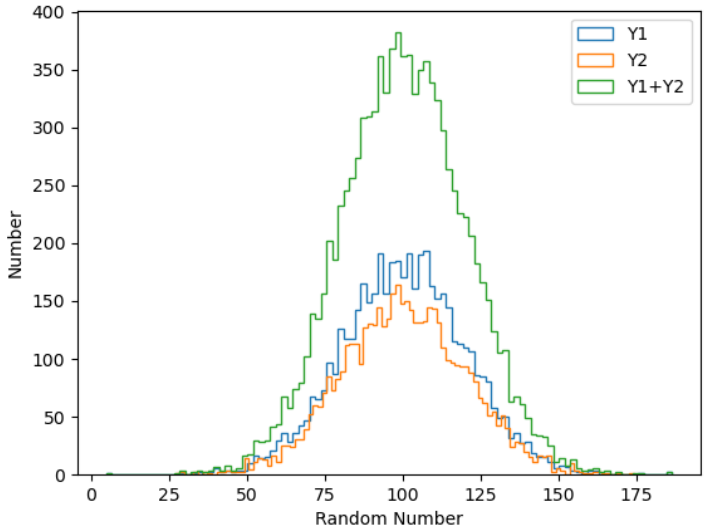


x_1, x_2은 균일한 분포를 가지지만
y_1, y_2는 Gaussian과 유사한 분포를 가진다.

x_1, x_2의 분포



y_1, y_2의 분포



토의사항

2.y_1, y_2사이에 상관관계가 존재하는가?

$$y_1 = \sqrt{-2\ln x_1}\cos 2\pi x_2$$

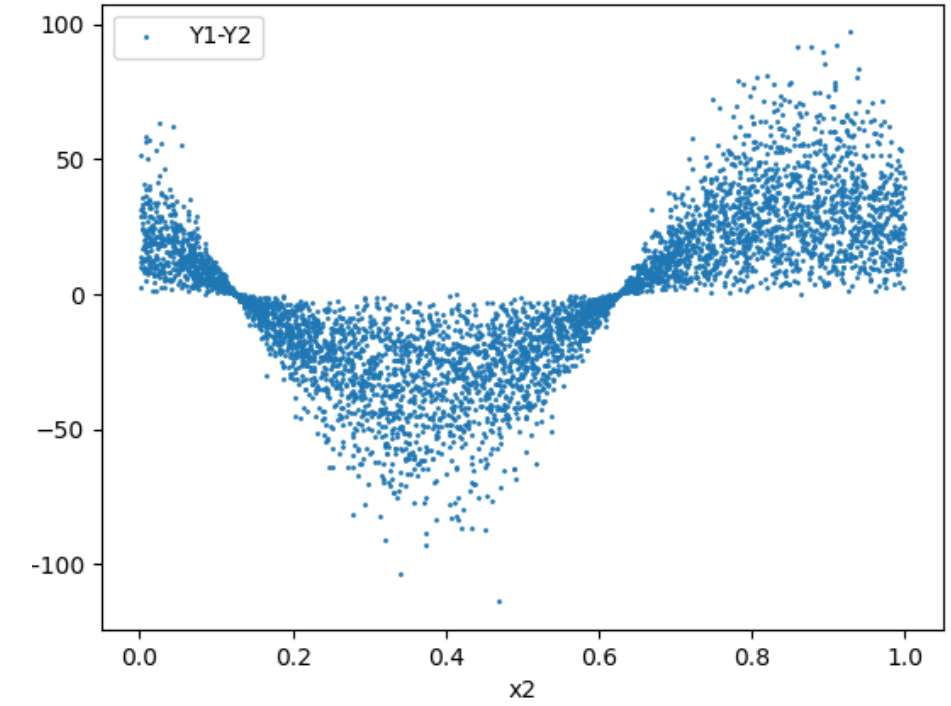
$$y_2 = \sqrt{-2\ln x_1}\sin 2\pi x_2$$

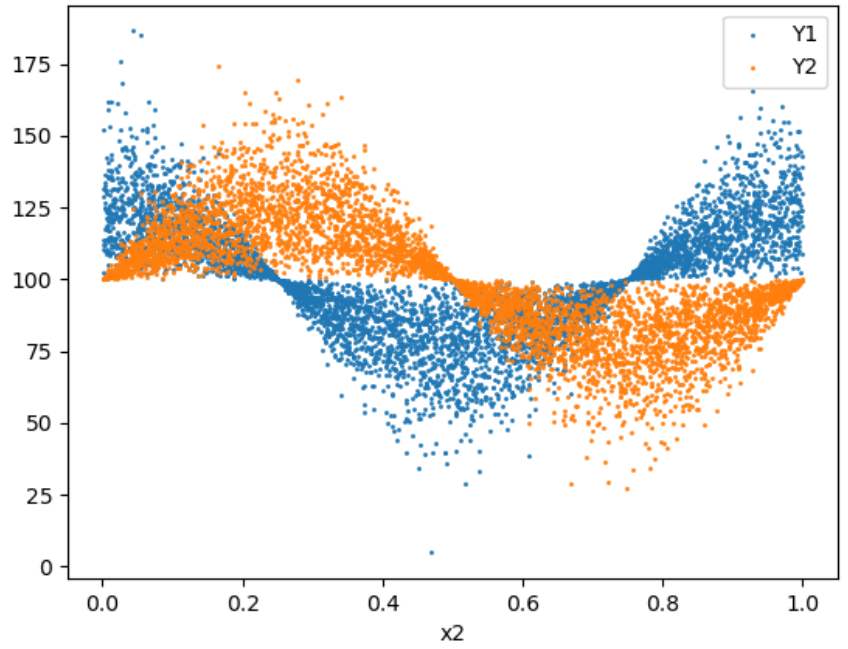
$$y_1 = \sqrt{-2\ln x_1}\sin(2\pi x_2 + \frac{\pi}{2})$$

$$y_2 = \sqrt{-2\ln x_1}\,sin\,2\pi x_2$$

로도 나타 낼 수 있다. -> $\frac{\pi}{2}$ 만큼 위상이동 했다고 볼 수있다.

y_1 과 y_2의 차이는 x_1의 값에 따라 그 크기가 바뀌지만 대략 삼각함수의 모습을 가진다.



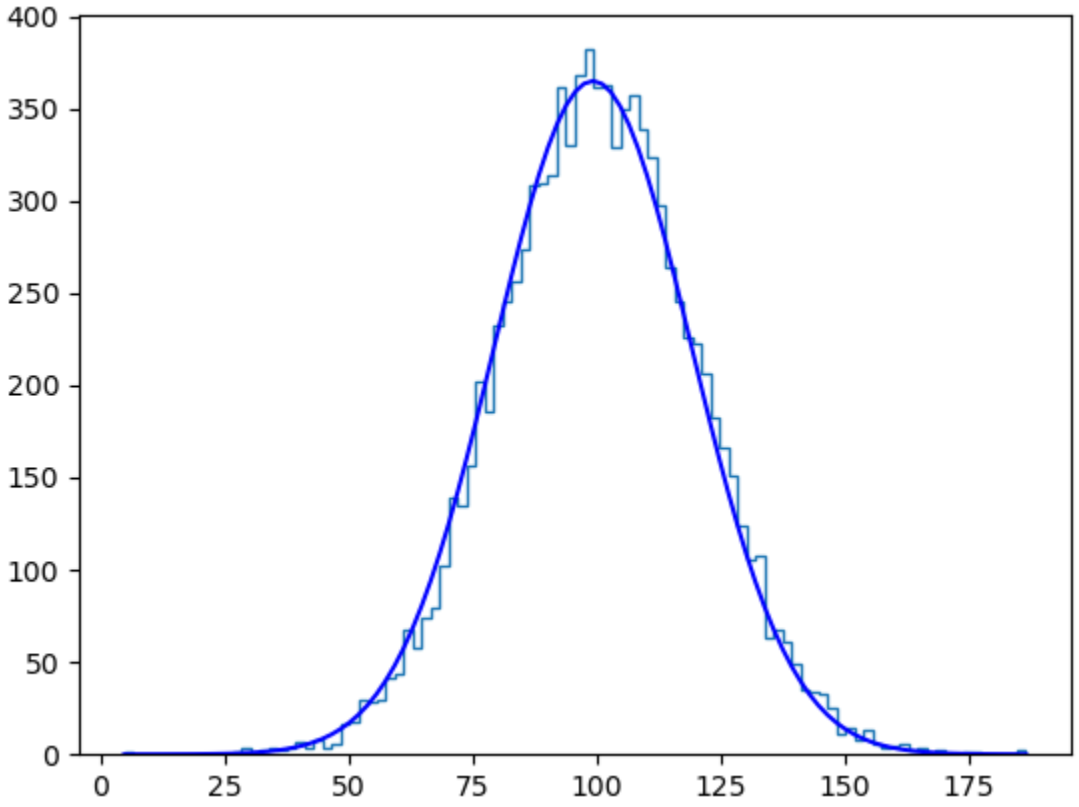


토의사항

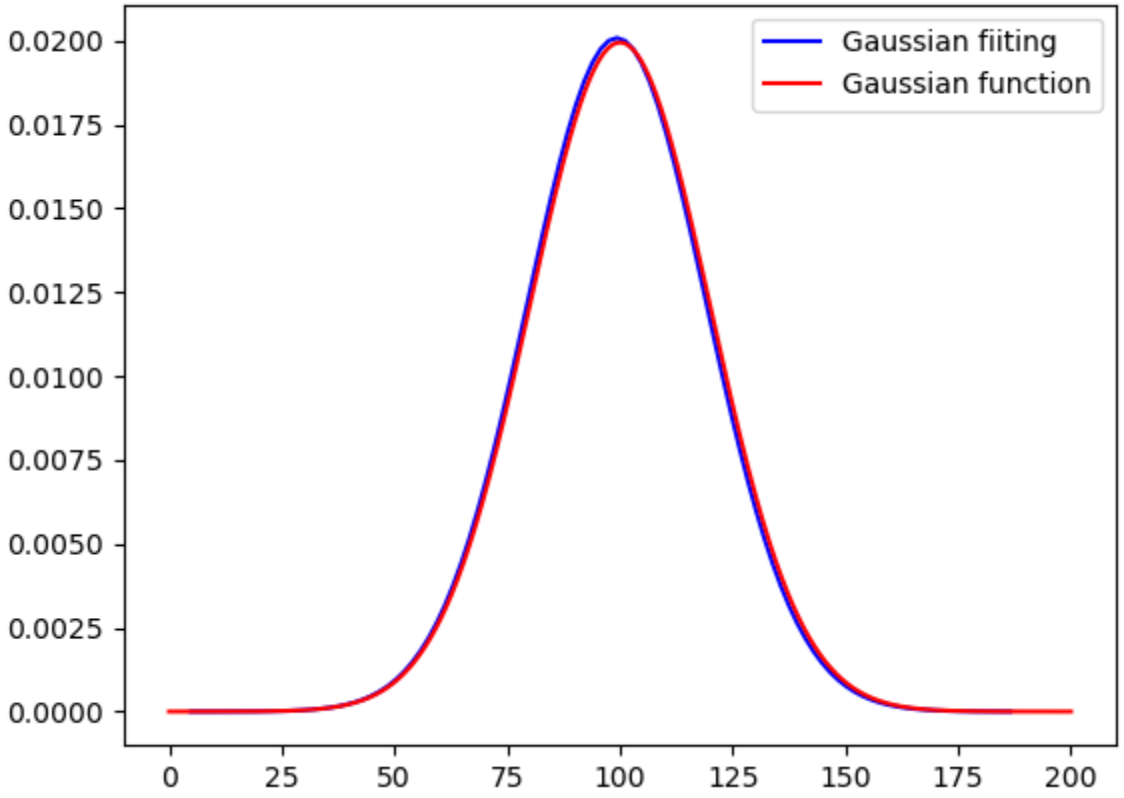
3.Gaussian 분포를 가지는가?

1)Curve_fit을 통한 Gaussian function fitting

Y3 Gaussian fitting 결과



평균 100, 표준편차 20을 가지는
Gaussian function의 확률밀도 함수와 비교



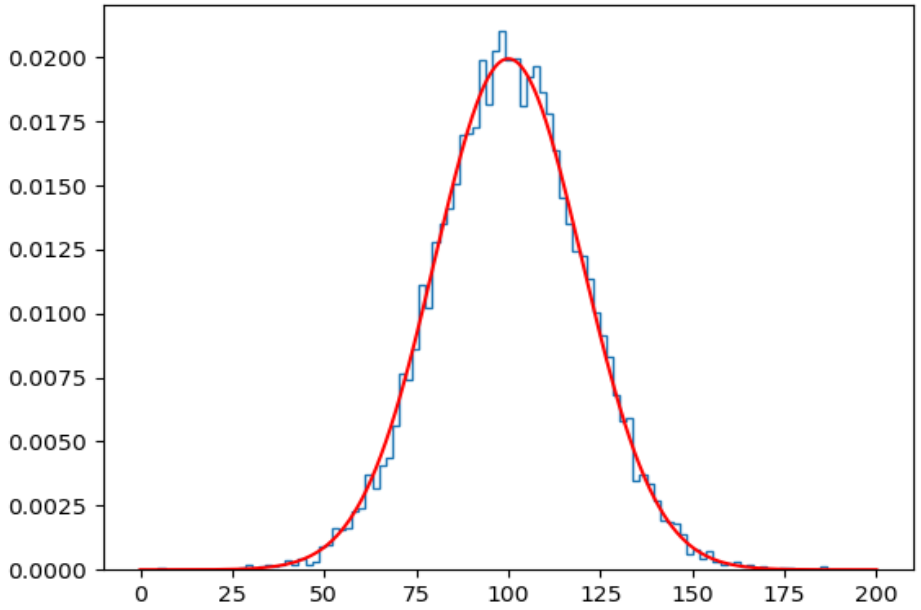
Gaussian function과 거의 일치하다
-> Gaussian이 제대로 fitting이 되었다

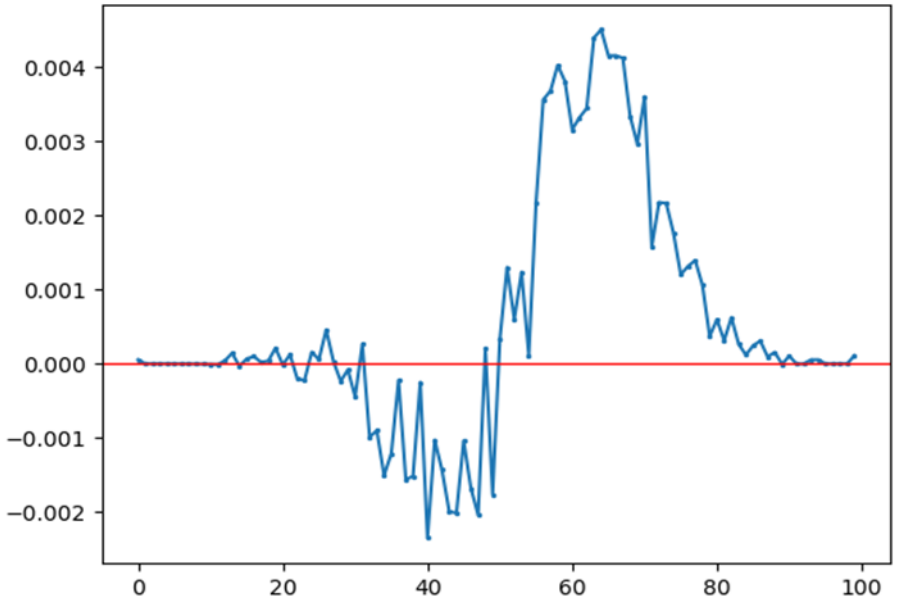
토의사항

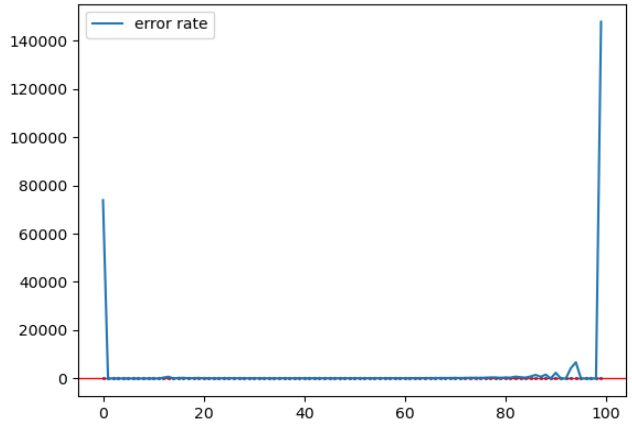
3.Gaussian 분포를 가지는가?

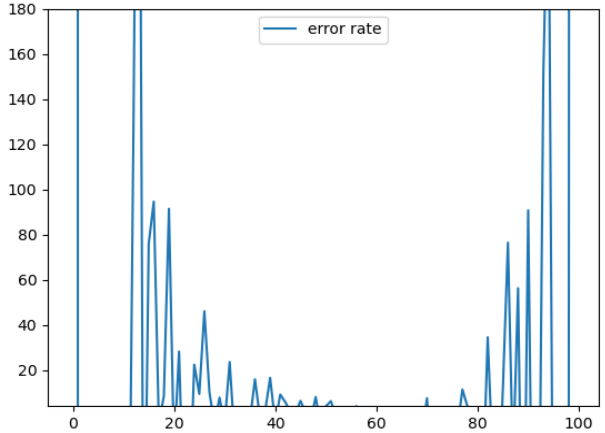
2)Gaussian function과 residual 비교

Gaussian function의 확률밀도 함수 값과 직접 만든 Gaussian 분포를 가지는 난수들의 확률 차이를 계산한다(bins=100 설정)









오차율 계산 시 중간으로 갈수록 0에 가까워진다.

토의사항

3.Gaussian 분포를 가지는가?

3-1)scipy.stats.normaltest (정규성 검정) 사용

P-value 값이 0.05가 넘으면 정규분포를 따른다고 볼 수 있다.

```
#정규성 검정 (normal test)
stats,p=stats.normaltest(mygauss(100,20,10000))
print(stats)
print(p)
```

N=10000인 경우 P : 0.05227201825686389
N=20000인 경우 P : 0.331170650098329
N=50000인 경우 P : 0.4392089894359118
N=100000인 경우 P : 0.5688812798165008

정규분포를 만족한다

3-2)scipy.stats.Anderson_ksamp (정규성 검정) 사용

Statistic 값이 cv의 값들보다 작으면 정규분포를 따른다.

```
#정규성 검정 (anderson)
statistic,cv,sl=stats.anderson(np.array(mygauss(100,20,50000)))
print(statistic)
print(cv)
```

N=10000인 경우

```
0.21331372550412198
[0.576 0.656 0.787 0.918 1.092]
```

N=50000인 경우

```
0.15093027543480275
[0.576 0.656 0.787 0.918 1.092]
```

정규분포를 만족한다

토의사항

3.Gaussian 분포를 가지는가?

4)Moment 계산

$$\mu_k = E[(X - \mu)^k]$$

Standardized Moments

$$\widetilde{\mu_k} = \frac{\mu_k}{\sigma^k}$$ σ = 표준편차

```
#moment 계산
X=np.array(mygauss(100,20,10000))
mean=stats.moment(X,1)/20**1
variance=stats.moment(X,2)/20**2
skewness=stats.moment(X,3)/20**3
kurtosis=stats.moment(X,4)/20**4
print(mean)
print(variance)
print(skewness)
print(kurtosis)
```

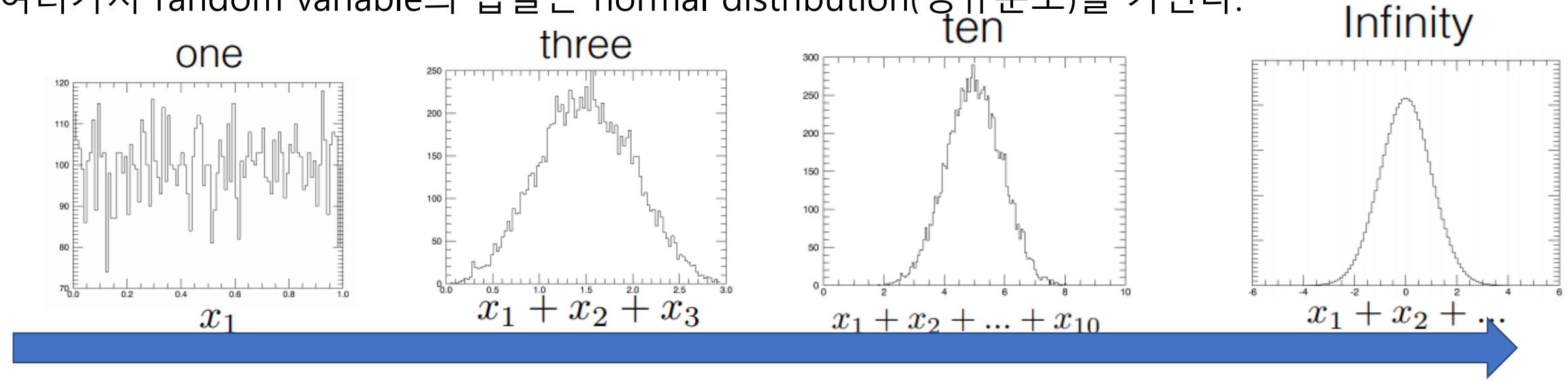
K=1 : 0.0
K=2 : 0.9974032709097349
K=3 : 0.002958420960836607 6
K=4 : 3.107919385580095

Standardized Moments인 경우
K=1 : 0.0
K=2 : 1
K=3 : 0
K=4 : 3

약간의 차이가 있지만 가우시안의 moment값과 유사하다

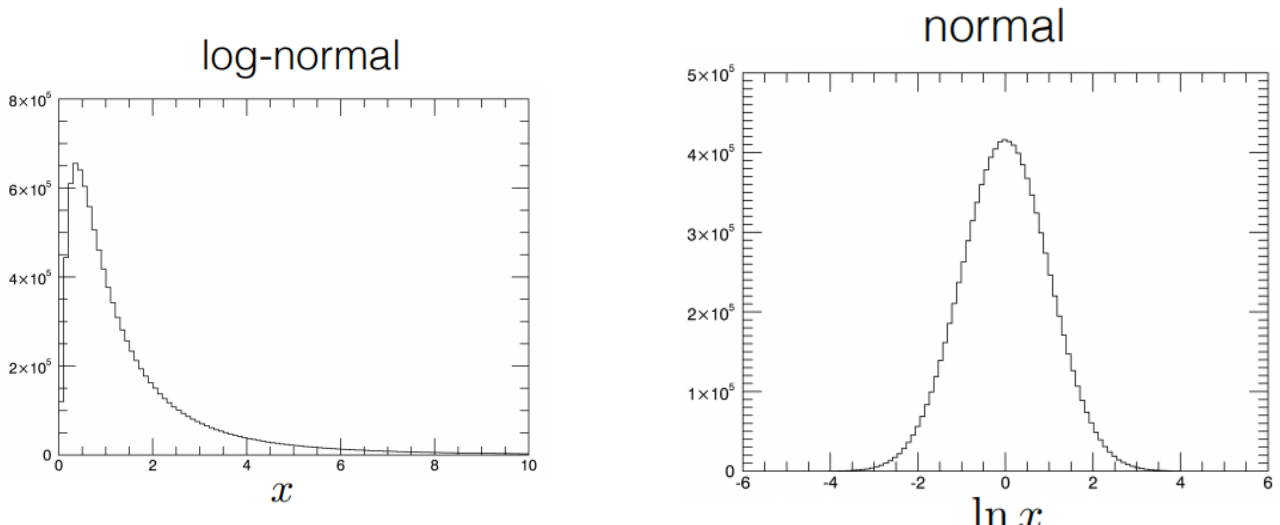
기본원리(Central Limit Theorem)

여러가지 random variable의 합들은 normal distribution(정규분포)를 가진다.



더하는 변수의 종류가 많아질수록 점점 더 정규분포를 가진다.

여러가지 random variable의 곱들은 log-normal distribution을 가진다



Log-normal distribution을 가지는 변수에 log를 취하면 정규분포를 가진다

실행

변수 10개 생성

```
a=np.array(myran(100000))
b=np.array(myran(100000))
c=np.array(myran(100000))
d=np.array(myran(100000))
e=np.array(myran(100000))
f=np.array(myran(100000))
g=np.array(myran(100000))
h=np.array(myran(100000))
i=np.array(myran(100000))
j=np.array(myran(100000))
```

```
ax1.hist(a, bins=100)
ax2.hist(a+b, bins=100)
ax3.hist(a+b+c, bins=100)
ax4.hist(a+b+c+d+e+f+g+h+i+j, bins=100)
```

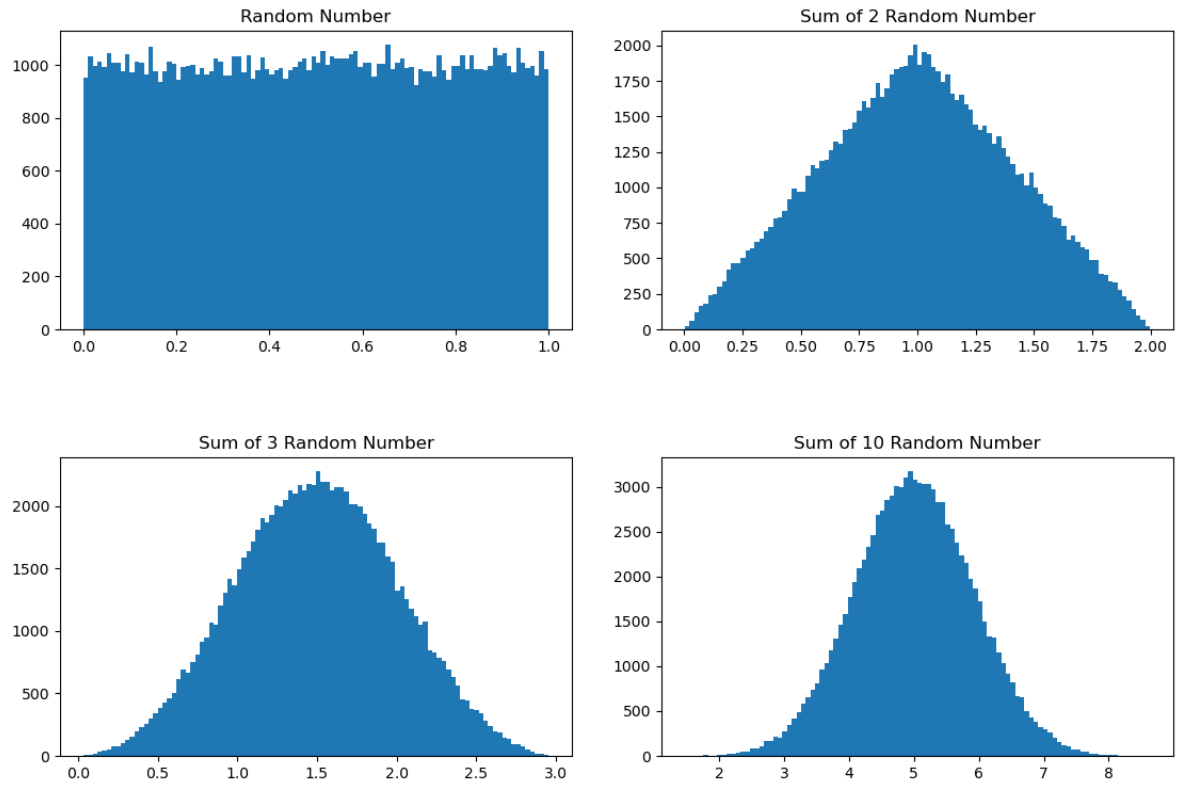
```
ax1.hist(a, bins=100)
ax2.hist(a*b, bins=100)
ax3.hist(a*b*c, bins=100)
ax4.hist(a*b*c*d*e*f*g*h*i*j, bins=100)
```

각각 변수 1개,2개,3개,10개 일때의
합과 곱들의 결과
히스토그램으로 표현

결과

난수들의 합 표현

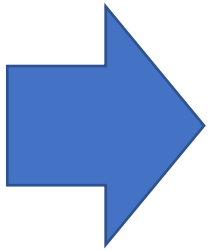
합하는 난수들이 많아질수록 정규분포의
형태를 가진다

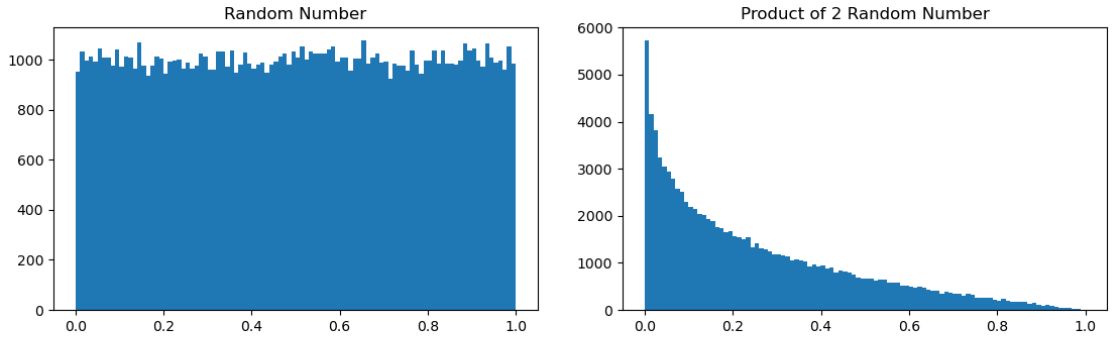


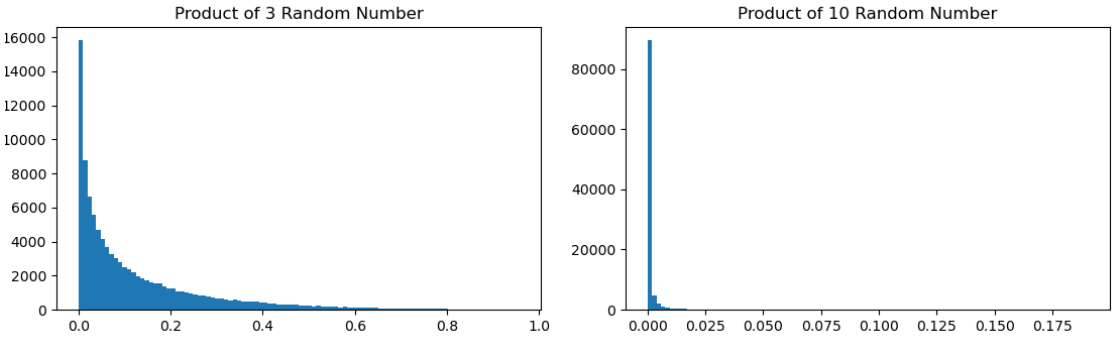
결과

난수들의 곱 표현

곱하는 난수들이 많아질수록 log normal distribution 의 형태를 가진다

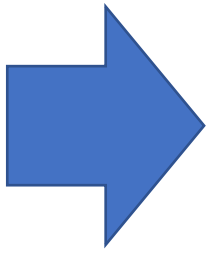


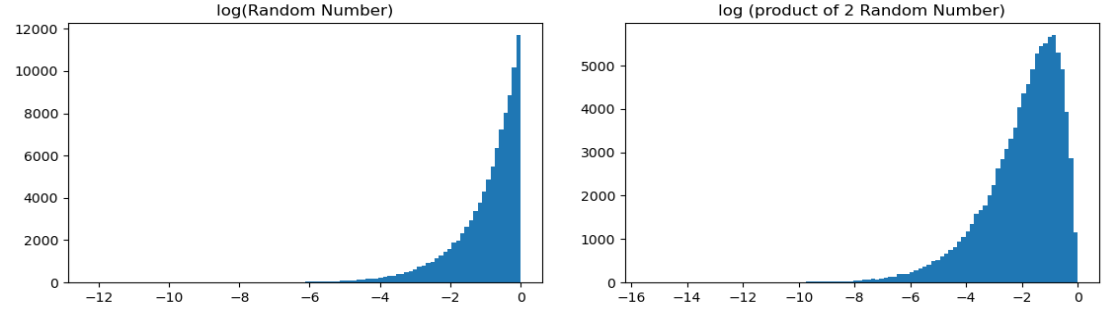


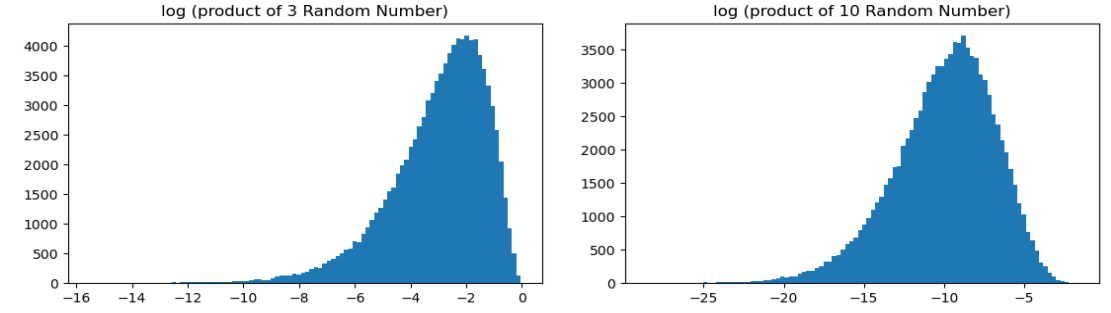


난수들의 곱의 log 표현

곱하는 난수들이 많아질수록 normal distribution 의 형태를 가진다

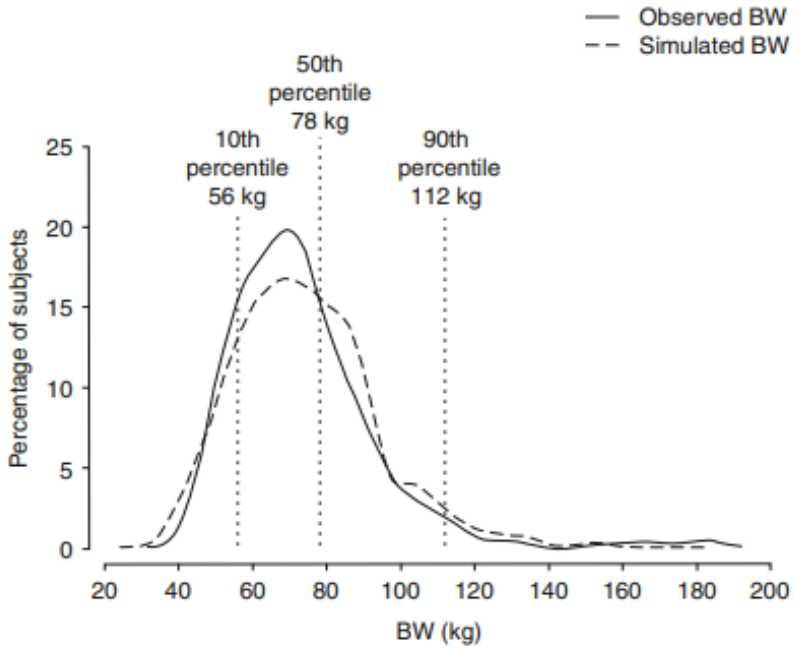






토의사항

키의 분포가 정규분포를 가짐을 확인하였다. 그렇다면 몸무게는?



Weight의 분포는 정규분포와 가까운 형태를 가진다.

키의 난수들의 합과 유사하게
1살대 몸무게 증가, 2살때 몸무게 증가,....n살대 몸무게 증가의 변수들의 합으로 나타낼수 있다.

혹은 여러 환경 요인들의 합(식습관,운동빈도,기초대사량등의 합으로도 나타낼수 있다.
-> Central Limit Theorem을 만족한다

Reference

1.Interpreting the Anderson darling test scipy,stackoverflow,2019/2/20

https://stackoverflow.com/questions/53909526/interpreting-the-anderson-darling-test-scipy

2.scipy.stats.moment ,scipy.org

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.moment.html?highlight=moments

3.scipy.stats.normaltest,scipy.org

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html

4.Plot a 3D bar histogram with python,stackoverflow,2018/09/19

https://stackoverflow.com/questions/52385299/plot-a-3d-bar-histogram-with-python

5.How to print Gaussian curve fitting results?,stackoverflow,2020/08/25

https://stackoverflow.com/questions/63585652/how-to-print-gaussian-curve-fitting-results