확률통계및프로그래밍 **Learning Journal**

Central Limit Theorem of the Stock Revenue

세종대학교 지능기전공학부 무인이동체공학전공
21011890 노지민

1. 서 론

1.1 Central Limit Theorem 의 중요성

현실에서 발생하는 자연현상을 확률분포로 나타내면, 대부분의 현상이 Gaussian RV 로 모형화가 가능하다. 이때 N 개의 임의의 분포로부터 얻은 표본의 평균은 N 이 증가할수록 기대값이 μ, 분산이 σ^2/N 인 Gaussian RV 로 수렴한다. 확률에서 이와 같이 정리할 수 있는 이유를 Central Limit Theorem 로 설명할 수 있기 때문에 Central Limit Theorem 은 확률론에서 매우 중요한 이론으로 손꼽힌다.

1.2 연구의 목적

본 저널에서는 확률론에서 중요한 Central Limit Theorem 을, 예시의 데이터를 바탕으로 Matlab 을 통한 시뮬레이션 방법을 통해 증명할 것이다. 이때 표본의 개수인 n 을 증가하며 진행했을 때의 시뮬레이션에서 두드러지는 특징에 주목하며, 보다 확실히 Central Limit Theorem 의 이해하고자 한다.

1.3 저널의 구성

본론에서는 확률 모델을 세우기 위한 주요 개념인 Random Variable, Continuous Random Variable, CDF and PDF, Gaussian RV, Central Limit Theorem 에 대해 정리하고, 최종적으로 Stock Revenue 의 예시에 대한 시뮬레이션으로 Central Limit Theorem 에 관해 증명하는 크게 두 가지 섹션으로 구성되어 있다.

2. 본 론

2.1 주요 개념 정리

본격적으로 확률 모델을 세우고 시뮬레이션하기 앞서 필요한 주요 개념들을 먼저 정리하고자 한다.

2.1.1 Random Variable

실제적인 관측을 통해 전개되는 확률 실험에서의 데이터를 바탕으로 모델링 한 확률 모델은, 확률 실험에서의 가능한 모든 관측의 집합인, Sample Space 의 원소에 실수로 대응시키는 함수로 볼 수 있는데, 이때의 함수를 Random Variable 이라고 한다. Random Variable 은 보통 대문자로 X 이와 같이 나타내고, 이때 X 가 가지는 모든 값들의 집합을 X 의 치역이다. 대부분의 경우 한 번에 여러 개에 해당하는 Random Variable 을 생각하기 때문에 Random Variable 의 치역을 S 로 표시하고, 해당하는 Random Variable 를 아래 첨자로 S_x 로 표시한다.

확률 모델은 위에서 정리한 것과 같이 항상 확률 실험으로부터 시작되기 때문에 각 Random Variable 은 직접적으로 확률 실험들과 상관관계가 있으며, 세 가지로 정리할 수 있다.

1) Random Variable 은 관측 그 자체이다.

Ex) 10^(-5)초 동안 도착하는 광전자의 수를 세는 경우, 우리는 각각의 관측을 Random Variable X 로 볼 수 있다. 그렇다면 X 의 치역은 s_x= {0,1,2, … }으로 정리할 수 있고 최종적으로 S 와 같으므로 Random Variable 은 관측 그 자체라고도 말할 수 있다.

2) Random Variable 은 관측의 함수이다.

Ex) 4 개로 이루어진 비트를 통해 데이터를 전송하는 패킷이 있다. 이때 각 비트가 제대로 동작할 때 1, 아닐 때 0 을 전송한다고 하자. 이때 0 혹은 1 이 4 개가 나열된 수열로 나타낼 수 있다. (s_3= 0010 과 같이 나타낼 수 있다.) Sample Space 는 총 16 가지의 가능한 수열들로 구성된다. N 을 제대로 동작하는 비트의 개수라고 한다면, N 은 확률 실험과 연관있는 Random Variable 이 된다. (s_3 의 결과에 대해서는 N = 1) 따라서 N 의 치역은 s_N

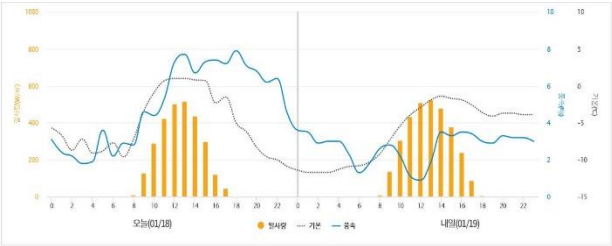
= {0,1,2,3,4} 이 된다. 즉, Random Variable N 은 관측, 확률 실험의 함수이다.

3) Random Variable 은 다른 Random Variable 의 함수이다.

Ex) 위의 예시에 이어 R 을 제대로 동작하지 않는 비트의 개수라고 정의한다면, Random Variable R 은 4-N 으로 N 에 대한 함수로 나타낼 수 있다. (s_3 의 결과에 대해서는 R = 3). 따라서 R 의 치역은 s_R = {4,3,2,1,0}으로 정의할 수 있다. 즉, Random Variable 은 다른 Random Variable 의 함수임을 알 수 있다.

위의 내용을 통해 Random Variable 의 값은 항상 해당 확률 실험의 결과로부터 도출할 수 있음을 알 수 있다. 하지만, 항상 확률실험의 결과값들에 숫자를 할당하는 과정과 관측이 같이 이루어지지는 않는다. 앞의 예제와 같이 확률 실험의 정의로부터 자연스럽게 정해지는 경우도 있고 추가적인 분석과 계산과정이 필요한 경우도 있다. 이러한 경우 확률 실험에서 명시된 관찰이 불가능한 Parameter 값도 존재할 수 있다. 이처럼 Parameter 을 포함하고 있는 확률 실험의 경우 Random Variable 를 만들어내지 않는다. 이와 같이 Random Variable 을 생성하지 못하는 Parameter 들을 갖는 확률 실험을 부적합한 확률실험이라고 한다.

그림 1. 부적합한 확률실험의 예시 - 시간대 별 일사량



(출처: SOLAR DIRECT, 전국 시간대별 발전량, 기상 예측 데이터)

시간대 별 일사량은, 항상 0~24 사이의 시간이라는 변수에 의해서만 정의될 수 없다. 해당 날의 명시된 관측이 불가능한 복합적인 자연 환경의 Parameter 를 포함하기 때문이다. 즉, 시간대 별 일사량은 부적합한 확률실험이며, Random Variable 을 만들어내지 못하는 예시에 해당한다.

위와 같이 Random Variable 과 확률 실험과의 연관성을 중심으로 Random Variable 에 대해 살펴보았다. Random Variable 는 가능한 값들을 셀 수 있는 집합으로 형성되는지, 셀 수 없는 집합으로 형성되는 지에 따라 이산적인 혹은 연속적인 Sample Space 를 갖는다. 각 경우를 Discrete Random Variable 과 Continuous Random Variable 로 나눌 수 있으며, 이에 따라 Random Variable 의 성질을 정리할 수 있다.

2.1.2 Continuous Random Variable

연속된 실수 집합을 치역으로 갖는 Random Variable 을 Continuous Random Variable 이라고 한다. 대표적인 Continuous Random Variable 에는 시간, 전압, 위상, 제품의 수명 등이 있다. 이 경우 어떤 정해진 구간 사이에 실선에 대응하는 무한히 많은 값들을 갖게 되는데, 이 값들 각각에 양의 확률을 부여하면 확률의 합은 더 이상 1 이 아니게 된다. 그러므로 Continuous Random Variable 에 대한 확률 분포를 생성하려면 Discrete Random Variable 에서와 다른 방식으로 접근해야 한다.

또한, Continuous Random Variable 에서는 각각의 한 점에 대한 결과의 확률이 0 이라는 점에 주의해야한다. 이 성질을 보다 확실히 이해하기 위해 한 가지 예를 들어 이해해보았다. 나무에 매달린 사과가 4 초동안 아래로 자유낙하 해 지면에 도달하는 상황을 뉴턴이 지켜보고 있다고 하자. 사과가 떨어지는 동안의 시간을 T 라고 할 때, T 를 Continuous Random Variable 로 모델링 할 수 있고, T 의 Sample Space 는 $s\_T=\{t|0\le t\le 4\}$ 라고 정리할 수 있다. 이때 속도가 아래방향으로 1m/s 에 도달하는 시간을 예측해 본다고 하자. 예측 구간이 작으면 작을수록 그 예측이 맞을 확률은 점점 작아지는데, 특정 속도인 1m/s 가 되기까지의 시간은 정해져 있으므로, 정확히 1m/s 에 도달할 때의 순간을 맞출 확률은 0 에 수렴한다. 이렇듯 Continuous Random Variable 이 어떤 구간에 속할 확률은 구간이 좁아지면 좁아질수록 점점 작아져 결국 Continuous Random Variable 에서의 한 점의 확률은 0 이며, 질량을 가지지 않는다.

이를 수학적으로 정리하면 아래와 같이 나타낼 수 있다. 시간 t 는 0~4 사이의 n 개의 X 를 가질 수 있고, n 의 개수는 무한하므로 한 점 x 에 대한 확률은 아래와 같이 나타낼 수 있다.

$$P[X=x] = \lim_{n\to\infty}(1/n) = 0$$

즉, 결과 x 에 상관없이 P[X = x] = 0 이다. 그러므로 Continuous Random Variable 에서 의미 있는 확률은, 구간에 부여되는 확률이라고 이해할 수 있다..

2.1.3 CDF and PDF

앞서 Continuous Random Variable 에서 한 점에 대해 확률은 0 으로, 질량을 가지지 않는다는 점을 강조했다. 이 개념은 물리학에서 연속된 부피의 질량을 생각하는 것과 유사하다. 물리학에서 한 물체가 일정 부피를 가지고 있어도, 한 점의 질량은 없는 것으로 보는데, 이런 경우 한 점에서도 해석을 하기 위해 밀도라는 개념을 이용한다. 확률론에서도 Continuous Random Variable 의 확률을 해석하기 위해 Probability Density Function 이라는 개념이 등장했다. PDF 에 대해 살펴보기 전에 Cumulative

Distribution Function 에 대한 개념을 알아야 보다 확실하게 개념을 이해할 수 있다.

CDF F_x(x)는 어떤 종류의 Random Variable 에 대해서도 확률 모델이 된다. 즉, Continuous Random Variable 에 대해 CDF 도 연속 함수가 되며 그 역도 성립한다. CDF 는 아래와 같이 정의된다.

$$F_x(x) = P[X \le x]$$

위의 정의에서 알 수 있듯이 CDF 그래프는 왼쪽에서 0 으로 시작하여 오른쪽에서 1 로 끝난다. 또한 그래프는 단조증가함수이다. 위의 정의에 의해 Random Variable 이 해당 구간에 속해 있을 확률은 그 구간의 양 끝점의 CDF 의 차이가 된다. 이 특성은 Random Variable 모든 X 에 대해 항상 성립한다. 위의 내용을 정리해 수식으로 나타내면 아래와 같다.

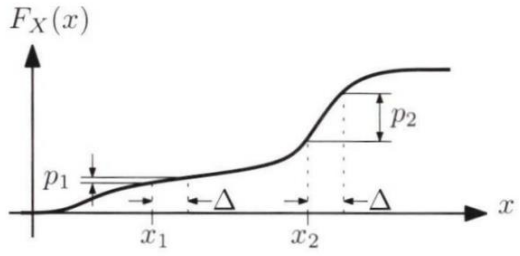
a) $F_x(-\infty) = 0$

b) $F_x(\infty) = 1$

c) $P[x_1 \le X \le x_2] = F_x(x_2) - F_x(x_1)$

위의 CDF 에 의해 Probability Density Function, PDF $f(x)$ 를 정의할 수 있다. CDF 의 경우처럼, PDF $f_x(x)$ 역시 Continuous Random Variable X 의 확률 모델이 된다. $f(x)$는 CDF 의 미분으로 정의할 수 있고, 이 값은 X 가 x 에 아주 가까이 있을 확률에 비례한다. CDF 의 미분 값은 CDF 그래프에서 해당 점의 기울기와 같은 값이므로 그래프를 통해 직관적으로 이해해보자.

그림 2. CDF F_x(x)의 그래프의 예



위의 그림과 같이 Random Variable X 가 $(x_1, x_1 + \Delta]$ 라는 너비 Δ인 구간에 속해 있을 확률 p_1과 X 가 $(x_2, x_2 + \Delta]$ 구간에 속해 있을 확률 p_2 는 아래와 같이 나타낼 수 있다.

$$p_1 = P[x_1 \le X \le x_1 + \Delta] = F_x(x_1 + \Delta) - F_x(x_1)$$
$$p_2 = P[x_2 \le X \le x_2 + \Delta] = F_x(x_2 + \Delta) - F_x(x_2)$$

두 경우 모두 구간 너비가 Δ인 경우인데, 이때 Δ 가 점점 작아진다면 p_1, p_2 에 대한 확률도 점점 작아져 각각 Random Variable X 가 x_1, x_2의 근처에 있을 확률이 된다. 이 값들은 각 점에서의 $F_x(x)$의 평균 기울기에 의해 정해진다.

$$P[x_1 \le X \le x_1 + \Delta] = \frac{F_x(x_1 + \Delta) - F_x(x_1)}{\Delta}\Delta$$

분수 형태로 되어 있는 우변의 식이 평균 기울기에 해당하며, 위 식은 Random Variable x_1 근처의 구간에 속할 확률은 평균 기울기에 해당 구간의 너비를 곱한 값이다. 이때 Δ -> 0 인 경우 평균 기울기는 $F_x(x)$의 x_1에서의 미분 값이 된다. 이렇게 해서 최종적으로 x 근처에서의 CDF 의 평균 기울기는 Random Variable X 를 x 근처에서 발견할 확률로 볼 수 있다는 결론에 도달할 수 있다. 물리학에서 아주 작은 부피의 물질의 질량이 그 물질의 밀도에 부피를 곱해 구해지는 것처럼, Continuous Random Variable 가 아주 작은 구간에 있을 확률은 CDF 의 기울기와 구간 너비의 곱으로 표현할 수 있으며, 이를 Probability Density Function 으로 정의할 수 있다.

Probability Density Function 이란 Continuous Random Variable X 에 따라 변하는 확률의 Density 를 뜻하며, PDF 는 아래와 같이 정의할 수 있다.

$$f_x(x) = \frac{d\,F_x(x)}{dx}$$

즉, 해당 구간에 대해 PDF 는 CDF 를 미분한 값이며, 반대로 CDF 는 PDF 를 적분한 값이다. 이에 PDF $f_x(x)$를 갖는 Continuous Random Variable X 에 대해 아래와 같이 정리할 수 있다.

a) $f_x(x) \ge 0 \;\; for\; all\; x$

b) $F_x(x) = \int_{-\infty}^{x} f_x(u)\,du$

c) $\int_{-\infty}^{\infty} f_x(x)\,dx = 1$

d) $P[x_1 \le X \le x_2] = \int_{x2}^{x1} f_x(x)\,dx$

2.1.4 Gaussian RV(μ , σ)의 개념 및 중요성

Continuous Random Variable 로 이루어진 연속 확률 분포는 종 형태, 지수함수 형태, 수평 형태 등 다양한 형태를 가질 수 있다. 그 중에서도 PDF 의 그래프가 종 모양의 곡선 형태를 띈 Gaussian Random Variable 은 실제 사건들의 그래프와 양상이 매우 비슷하고, 평균값인 μ 와, 분산 값인 σ^2 값만 안다면 바로 PDF 값과 그래프 형태도 잡을 수 있기 때문에 매우 중요한 Random Variable 로 손꼽힌다. Gaussian RV 는 아래와 같이 PDF 가 정의된다.

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

겉보기에는 굉장히 복잡한 형태를 띄고 있지만, 평균값인 μ 와, 분산 값인 σ^2 값만 안다면 바로 PDF 값과 그래프를 구할 수 있기 때문에 실생활에 많이 응용된다. 종의 중심은 μ이고, σ은 종의 너비를 알 수 있다. 더 나아가 σ값이 상대적으로 작다면 종 모양의 PDF 그래프는 더 높고 뾰족한 꼭대기를 가지며, 폭은 좁을 것이다. 반면에 σ가 크면, PDF 그래프는 상대적으로 낮고 평평한 꼭대기와 넓은 폭을 갖는다. 또, PDF 그래프는 평균 μ에 대해 좌우대칭이다. PDF 정리에 의해 Gaussian RV 에서의 PDF 그래프의 총 면적은 1 과 같으므로 평균 값의 왼쪽과 오른쪽 면적은 0.5 로 같음을 볼 수 있다. 위의 내용을 정리하자면, X 가 Gaussian RV 일 때, E[X] 와 Var[X] 값은 아래와 같다.

$$E[X] = \mu$$

$$Var[X] = \sigma^2$$

그렇다면, Gaussian RV 에서 CDF 는 어떻게 구할 수 있을까? 앞서 Gaussian RV 의 PDF 식을 통해 CDF 를 구하기란 불가능하다. 그렇기 때문에 Gaussian RV 의 주요 속성을 통해 다른 방법으로 CDF 에 접근해보고자 한다. Gaussian RV 에서 중요한 속성은 아래와 같다.

X 가 Gaussian(μ, σ) 일 때, Y = aX + b 는 Gaussian RV$(a\mu + b, a\sigma)$ 이다.

위의 속성을 통해 Gaussian RV 의 모든 선형 변환이 또 다른 Gaussian RV 를 생성한다는 점을 알 수 있다. 즉, 임의의 Gaussian RV 의 속성을 어느 특정한 Gaussian RV 의 속성과 연관해 이해할 수 있도록 한다.

위의 속성을 사용해 Gaussian RV(0,1)의 값을 갖는 경우를 Standard Normal Random Variable Z 라고 한다. 이 경우 $E[X] = 0,\ Var[X] = 1$인 Gaussian RV 이다. 이때, μ, σ의 값이 정해졌고, PDF 식에 대입했을 때 식이 정리가 되므로, 적분을 통해 CDF 를 구할 수 있게 된다. 이때의 Z 의 CDF 인 $F_z(z)$를 아래와 같이 나타낼 수 있다.

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-u^2/2}\,du$$

위 식은 x 가 X 의 기댓값으로부터 표준편차의 몇 배만큼 떨어져 있는지 판단할 수 있는 척도인 z-score 와 공식이 같음도 알 수 있다. 이제 위 식을 z 자리에 대입해 X 가 Gaussian RV(μ, σ)일 때의 CDF 와 확률을 정리해보자. 정리하면 아래와 같이 나타낼 수 있다.

$$F_x(x) = \Phi(\frac{x-\mu}{\sigma})$$

$$P[a < X \le b] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi(\frac{a-\mu}{\sigma})$$

2.1.5 Central Limit Theorem

현실에서 발생하는 많은 사건들은 Gaussian RV 로 근사한다. 이것은 Central Limit Theorem 에 의해 설명할 수 있다. Central Limit Theorem 은 모집단에서 추출된 측정값들의 여러 확률 변수의 합이 근사적으로 Gaussian RV 를 따른다는 개념이다. Central Limit Theorem 은 유한인 평균 μ와 표준편차 σ를 가지면서 Gaussian 을 따르지 않는 모집단에서도 n 개의 랜덤한 표본이 추출되었을 때, n 이 충분히 크다면, 표본 평균 $\overline{X}$ 의 표집분포는 근사적으로 Gaussian RV 이며 평균 μ와 표준편차 $\sigma/\sqrt{n}$ 을 가진다. 이 근사는 n 이 커질수록 더욱 더 정확해진다.

이를 수식으로 전개해보도록 하자. 우선 모집단으로부터 추출된 임의의 N 개의 표본에 대한 Sample mean 은 아래와 같이 나타낼 수 있다.

$$\overline{X_N} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

이때 $\overline{X_N}$의 기댓값은 다음과 같다.

$$E[\overline{X_N}] = E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[X_i]$$
$$= \frac{1}{N}(N \times \mu) = \mu$$

또한, $\overline{X_N}$의 분산값은 다음과 같이 전개할 수 있다.

$$Var[\overline{X_N}] = Var\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N^2}\sum_{i=1}^{N} Var[X_i]$$
$$= \frac{1}{N^2}(N\sigma^2) = \frac{\sigma^2}{N}$$

Central Limit Theorem 은 무엇보다 통계적 추론에 있어 중요하다. 그 이유는 모집단의 모수들에 대한 추론을 하기 위해 많은 통계량들의 표본 측정치들의 합이나 평균들을 사용하는데, 이때 표본의 크기가 충분히 크다면 측정치들은 근사적으로 Gaussian RV 의 표집 분포를 갖다고도 생각할 수 있다. 그러므로 추론에 있어 Gaussian RV 를 사용하여 얻은 통계량들의 성질을 설명하고 표본 결과에 대한 확률을 계산할 수 있다.

2.2 Central Limit Theorem of the Stock Revenue Simulation

그렇다면 본격적으로 Central Limit Theorem 을 실제 데이터를 바탕으로 Matlab 으로 시뮬레이션하여 증명해볼 것이다. 이때 표본의 개수인 n 을 증가하며 진행했을 때, Gaussian RV 로 근사하는지 살펴보자.

Example of the Stock Revenue

단기투자를 목표로 (주)삼성 전자의 주식을 11 월 16 일 종가의 가격으로 매입하여 11 월 17 일 하루 중 랜덤한 시간 때에 매매한 사람들 8000 명을 대상으로 수익을 Random Variable 로 설정해 조사하려고 한다. 주식을 판 사람들의 하루 동안의 수익(R)은 아래의 식과 같이 구할 수 있다.

수익(R) = 매매 가격 - 매입 가격

이때, 매매 시간은 랜덤하게 정한다고 가정해, 사람들 간 주식을 매매한 가격의 확률 분포는 Uniform 하다고 가정했다.

11 월 16 일 기준 (주)삼성 전자의 한 주당 매입 가격은 종가인 63,000 원이고, 11 월 17 일 하루 동안의 (주)삼성 전자의 저가는 60,000 원, 고가는 68,000 원이다. 모집단에서 10 명의 사람들을 랜덤하게 뽑아 수익(R)을 구해 평균을 구하는 것을 시행 n=1 이라 했을 때, 시행 횟수 n 을 늘려 Central Limit Theorem 이 성립함을 보여라.

주식의 저가, 고가, 매입가 설정

```
minStock = 60000;
maxStock = 68000;
buyStock = 63000;
```

시행의 횟수 n = 10 일 때

```
% Define implement times
n=10;


% Initialize Seed for Rand
rng(0,'twister');


% Initialize variable of R and mean
R10 = zeros(1,10);
meanX10= zeros(1,10);


for i=1:n
    % Define Sell_Price (minStock ~
maxStock)
    Sell_Price10= (maxStock-
minStock).*rand(10,1)+minStock;
    % Revenue = Sell_Price - buyStock
    R10 = Sell_Price10 - buyStock;
    % Define mean of each implement
    meanX10(1,i) = mean(R10);
end


meanX10
```

meanX10 = 1×10

103 ×

1.9908 2.2817 1.6985 0.1825 1.4909 0.6000 1.1234 0.9753 1.7838 0.1127

```
histogram(meanX10)
title('10 명의 수익의 평균 X 를 10 번 수행
시 Outcome');
xlabel('수익의 평균 X (단위: 원)');
```

```
ylabel('같은 구간의 X 값이 나오는 횟수
(단위: 회)');
grid on
```

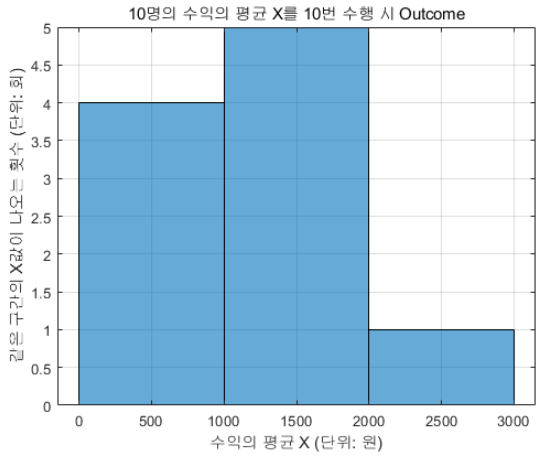


그림 3.
10 명의 수익의 평균 X 를 10 번 수행 시 Outcome

```
mean(meanX10)
```
```
ans = 1.2240e+03
```
```
var(meanX10)
```
```
ans = 5.6637e+05
```

시행의 횟수 n = 100 일 때

```
% Define implement times
n=100;


% Initialize Seed for Rand
rng(0,'twister');


% Initialize variable of R and mean
R100 = zeros(1,100);
meanX100= zeros(1,100);


for i=1:n
    % Define Sell_Price (minStock ~
maxStock)
    Sell_Price100= (maxStock-
minStock).*rand(10,1)+minStock;
    % Revenue = Sell_Price - buyStock
    R100 = Sell_Price100 - buyStock;
    % Define mean of each implement
    meanX100(1,i) = mean(R100);
end


meanX100
```

meanX100 = 1×100

103 ×

1.9908	2.2817	1.6985	0.1825	1.4909
0.6000	1.1234	0.9753	1.7838	0.1127
0.9346	0.7682	1.0626	0.3705	-0.1167
1.0102	0.6618	0.2928	0.6022	1.8101
2.0341	-0.3524	0.9607	0.2462	0.6805
1.6445	0.8927	1.6193	1.9028	0.7270
0.1855	1.4229	0.9098	-0.2830	1.9967
1.0392	1.1268	0.2209	0.9981	0.6476
0.2344	1.5384	0.4354	1.5583	0.8296
0.7195	0.6502	0.5825	1.1911	1.6913

```
histogram(meanX100)
title('10 명의 수익의 평균 X 를 100 번 수행
시 Outcome');
xlabel('수익의 평균 X (단위: 원)');
ylabel('같은 구간의 X 값이 나오는 횟수
(단위: 회)');
grid on
```

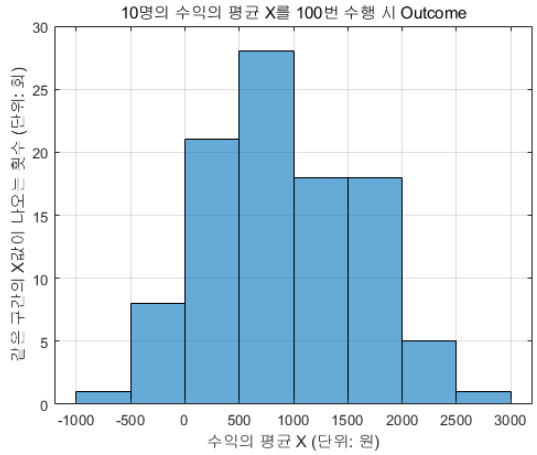


그림 4.
10 명의 수익의 평균 X 를 100 번 수행 시 Outcome

```
mean(meanX100)
```
```
ans = 910.6609
```
```
var(meanX100)
```
```
ans = 5.2493e+05
```

시행의 횟수 n = 1000 일 때

```
% Define implement times
n=1000;


% Initialize Seed for Rand
rng(0,'twister');


% Initialize variable of R and mean
R1000 = zeros(1,1000);
```

```
meanX1000= zeros(1,1000);


for i=1:n
    % Define Sell_Price (minStock ~
maxStock)
    Sell_Price1000= (maxStock-
minStock).*rand(10,1)+minStock;
    % Revenue = Sell_Price - buyStock
    R1000 = Sell_Price1000 - buyStock;
    % Define mean of each implement
    meanX1000(1,i) = mean(R1000);
end
meanX1000
```

meanX1000 = 1×1000

103 ×

1.9908	2.2817	1.6985	0.1825	1.4909
0.6000	1.1234	0.9753	1.7838	0.1127
0.9346	0.7682	1.0626	0.3705	-0.1167
1.0102	0.6618	0.2928	0.6022	1.8101
2.0341	-0.3524	0.9607	0.2462	0.6805
1.6445	0.8927	1.6193	1.9028	0.7270
0.1855	1.4229	0.9098	-0.2830	1.9967
1.0392	1.1268	0.2209	0.9981	0.6476
0.2344	1.5384	0.4354	1.5583	0.8296
0.7195	0.6502	0.5825	1.1911	1.6913

```
histogram(meanX1000)
title('10 명의 수익의 평균 X 를 1000 번 수행
시 Outcome');
xlabel('수익의 평균 X (단위: 원)');
ylabel('같은 구간의 X 값이 나오는 횟수
(단위: 회)');
grid on
```

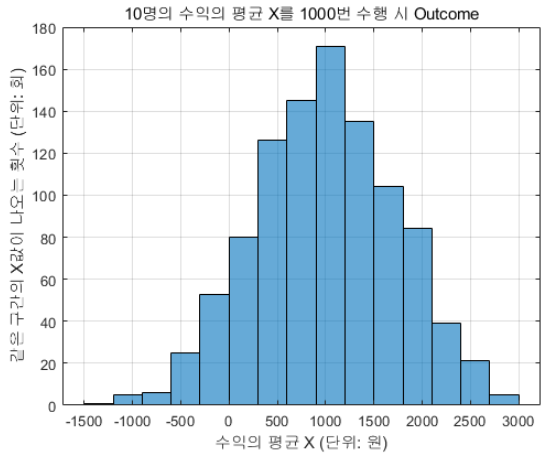


그림 5.
10 명의 수익의 평균 X 를 1000 번 수행 시 Outcome

```
mean(meanX1000)
ans = 996.4846
```

```
var(meanX1000)
ans = 5.3016e+05
```

시행의 횟수 n = 100000 일 때

```
% Define implement times
n=100000;


% Initialize Seed for Rand
rng(0,'twister');


% Initialize variable of R and mean
R100000 = zeros(1,100000);
meanX100000= zeros(1,100000);


for i=1:n
    % Define Sell_Price (minStock ~
maxStock)
    Sell_Price100000= (maxStock-
minStock).*rand(10,1)+minStock;
    % Revenue = Sell_Price - buyStock
    R100000 = Sell_Price100000 -
buyStock;
    % Define mean of each implement
    meanX100000(1,i) = mean(R100000);
end
meanX100000
```

meanX100000 = 1×100000

103 ×

1.9908	2.2817	1.6985	0.1825	1.4909
0.6000	1.1234	0.9753	1.7838	0.1127
0.9346	0.7682	1.0626	0.3705	-0.1167
1.0102	0.6618	0.2928	0.6022	1.8101
2.0341	-0.3524	0.9607	0.2462	0.6805
1.6445	0.8927	1.6193	1.9028	0.7270
0.1855	1.4229	0.9098	-0.2830	1.9967
1.0392	1.1268	0.2209	0.9981	0.6476
0.2344	1.5384	0.4354	1.5583	0.8296
0.7195	0.6502	0.5825	1.1911	1.6913

```
histogram(meanX100000)
title('10 명의 수익의 평균 X 를 100000 번
수행 시 Outcome');
xlabel('수익의 평균 X (단위: 원)');
ylabel('같은 구간의 X 값이 나오는 횟수
(단위: 회)');
grid on
```

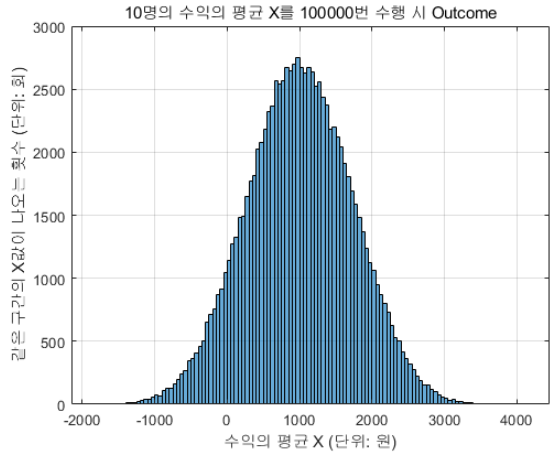


그림 6.
10 명의 수익의 평균 X 를 10000 번 수행 시 Outcome

```
mean(meanX100000)
ans = 1.0026e+03
var(meanX100000)
ans = 5.3143e+05
```

각 시행의 크기에 따라 for 문을 반복하여, rand() 함수를 사용하여 랜덤하게 매매 가격이 정해지는 10 명에 대한 수익(매매가격 - 매입가격)의 평균을 mean() 함수로 구해 meanX 변수에 저장했다. 최종적으로 위와 같이 Matlab 의 histogram() 함수를 사용해 meanX 의 그래프를 그렸다.

3. 결론

지금까지 Central Limit Theorem 을 증명하기 위해 Random Variable, PDF, CDF, Gaussian RV 등의 개념을 알아보고, 직접 예시를 들어 시뮬레이션을 통해 CLT 가 성립함을 증명했다.

앞서 예시를 보다 구체적으로 해석하자면, n=1 일 때, R 은 -3000 에서 +5000 까지의 범위를 갖고, 구간에서 모두 동일한 P(x)를 갖는다. 그러므로 모집단에서의 평균 μ 는 중간 값인 1000 을 갖는다. 그렇다면, n 의 시행을 계속해 늘리며 표본에 대한 그래프를 그렸을 때에도 모평균 μ=1000 값에 근사해야 Central Limit Theorem 이 성립한다. 그렇다면 실제로 시뮬레이션 했을 때의 결과값을 분석해보자.

첫 번째로 그림 4, 5, 6, 7 을 참고해, n 에 따른 그래프의 형태에 주목해보자. n 이 증가할수록 더욱 뚜렷한 종 모양을 띄고, 평균값을 기준으로의 대칭성도 높아진다. 즉, Gaussian RV 에 근사하다는 것을 볼 수 있다.

두 번째로 평균값에 주목해 보자. 앞서 모집단에서의 평균은 μ=1000 임을 알았다. 그렇다면 위의 표본들에 대한 평균은 어떤 값을 특징을 갖는지 살펴보자. 각 평균 값은 Matlab 의 Mean()함수를 사용해 구할 수 있다. 정리하면 아래와 같다.

$n=10,\ E[\bar{x}] = 1224$

$n=100,\ E[\bar{x}] = 910.6609$

$n=1000,\ E[\bar{x}] = 996.4846$

$n=100000,\ E[\bar{x}] = 1002.6$

n 이 증가함에 따라 표본의 평균값은 모집단의 평균 μ=1000 값과 오차가 줄어들며 근사함을 알 수 있다.

즉, uniform 했던 모집단에서, 충분히 큰 n 개의 관측치를 가진 랜덤 표본이 추출되었을 때, 표본 평균 $\overline{X}$의 표집 분포는 근사적으로 Gaussian RV 를 따르며, 평균 μ 를 가지므로 Central Limit Theorem 이 성립함을 증명할 수 있다.

참고문헌

1) 홍종선, 권태완,
[수익률 분포의 적합과 리스크값 추정] (2010.3)
2) 김규형, 이준행,
[극한치이론을 이용한 VAR 추정치의 유용성과 한계 - 우리나라 주식시장을 중심으로] (2005.6)
3) 장연홍 (2002.3),
[고전확률론과 중심극한정리에 대한 역사적 고찰]
4) Roy D.Yates, David J. Goodman,
[Probability and Stochastic Processes 3rd edition]
5) William Mendenhall, Robert J. Beaver, Barbara M. Beaver,
[Introduction to Probability & Statistics 14th edition].