확률과 통계 Assignment1

이름 : 이준휘

학번 : 2018202046

교수 : 심동규 교수님

강의 시간 : 월 수

1. Introduction

해당 과제는 Uniform distribution의 random variable X를 변환하여 Gaussian distribution을 띄는 random variable Y를 만드는 것이다. 이를 통해 만들어진 변수는 검증을 통해 해당 결과가 얼마나 유의한가에 대해 확인한다.

1. Approach

기본적으로 Gaussian distribution의 공식은 이다. 우리가 원하는 것은 Random Value X가 0~1 구간으로 주어졌을 때 나오는 Y가 Gaussian distribution을 따르도록 해야 하는 것이다. 기본적으로 Gaussian distribution의 CDF인 F(x)에서 이다. 또한 ) = 0으로 정의되어 있다. 이를 통해 우리는 CDF의 결과가 0~1구간으로 나오는 것을 알 수 있다. 이를 반대로 말한다면 CDF의 역함수를 통해 우리는 Gaussian distribution의 분포를 띄는 Y를 만들 수 있다는 것이다. 하지만 기본적으로 Gaussian distribution의 CDF도 일반적으로 계산하기 어려워 평균 0, 분산 1인 표준분포를 변환하여 이를 관찰하고 있다. 그렇기 때문에 이에 적분의 역함수를 구한다는 것은 어려운 일이다. 하지만 다음 방법을 사용하면 Gaussian distribution을 구하는 것이 가능하다.

필자가 찾은 방법은 Box-Muller transform(박스-뮬러 변환)를 이용한 Marsaglia polar method(극좌표법)이다. 해당 이론에 앞서 박스-뮬러 변환에 대해 설명하겠다.

박스-뮬러 변환이란 정규 난수 발생법 중 하나로 1958년에 발표되었다. 해당 변환에서는 기본적으로 2개의 Random Value를 사용한다. 로 가정하였을 때 를 좌표계에 넣고 극좌표로 나타내면 가 되고 으로 나타낼 수 있다. 여기서 Gaussian distribution을 이용하면 로 나타낼 수 있다. 또한 이의 CDF인 이다 여기서 치환 적분법을 사용하면 로 나타낼 수 있다. 여기서 지수분포인 F(x)를 살펴보자. 구간에서 의 값을 나타낸다. 여기서 위의 계산 결과와 지수분포가 유사한 것을 할 수 있다. 즉 을 띈다는 것을 알 수 있다. 기본적으로 이기 때문에 (x < 0)인 경우가 발생하지 않는다. 해당 결과에서의 구간은 임으로 구간을 바꾸기 위해 log값을 취해 변환하면 -2logU로 바꿀 수 있다. θ의 경우 의 범위에서 랜덤하게 선택하면 된다. 따라서 박스-뮬러 변환법에 의해 로 나타낼 수 있는 것이다. 또한 이 둘은 N(0, 1)을 성립한다.

극좌표법은 해당 공식에서 복잡한 연산이 많이 쓰인다는 단점을 고안한 방법이다. 박스-밀러 변환에서 cos과 sin으로 나타내였던 것을 없애고 대신에 (-1~1) 좌표 v1과 v2를 얻으며 해당 좌표의 길이가 1이하인 경우에만 난수를 발생시킨다. 이를 식으로 나타낼 경우 로 나타낼 수 있다. 를 해야한다.

필자의 사용 언어는 C++이며 g++환경에서 코드를 작동시켰다.

아래는 Gaussian random value를 생성하는 함수다.

텍스트이(가) 표시된 사진

자동 생성된 설명

해당 코드를 살펴보면 x1과 x2를 통해 -1 ~ 1범위의 double형 난수를 얻으며 해당 좌표의 길이의 제곱값을 pw에 저장한다. 만약 pw의 길이가 1을 넘어설 경우 다시 난수를 받아오고, 아닐 경우 위쪽에서 계산된 극좌표법의 식에 의거하여 값을 반환한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

위의 함수는 Histogram을 출력하는 함수다. 해당 함수는 값을 정렬하고 21개의 구간으로 분리한 뒤 각각의 범위에서의 개수를 출력해준다.

텍스트이(가) 표시된 사진

자동 생성된 설명

메인 함수는 다음과 같이 구성되어있다. 난수의 개수는 100개로 설정되어있다. 100개의 value를 저장할 배열과 입력값을 저장할 변수들이 설정되어있다. 입력을 통해 평균과 분산을 입력받으며 분산값을 통해 표준편차를 저장해둔다. 그 후 for문을 통해 이전에 생성한 Gaussian\_distribution함수를 통해 난수를 생성하며 이를 출력한다. 출력한 난수는 Print\_Histogram함수를 통해 정렬, 히스토그램 출력이 된다. 이후에는 검증구간이다.

1. Verification

필자의 검증방식은 chi-Square Verification(카이 제곱 검증)이다. 서로 독립인 이 표준정규분포 N(0,1)을 따른다면 이라 하고 이때의 를 자유도가 v인 카이제곱 확률변수라고 말한다. 이를 정규 분포 에서로 바꾸게 되면 으로 나타낼 수 있다. 여기에서 X’와 s^2(표본 분산)이 독립이고 일 때 인 것이다. 또한 이를 0.5, 0.1, 0.05, 0.01의 값을 가진 P\_value(진위의 가설에서 희소나 극한으로 값을 얻을 확률)에 해당하는 값과 비교한다면 해당 난수의 정확성을 알 수 있다.

텍스트, 스크린샷, 화면, 닫기이(가) 표시된 사진

자동 생성된 설명

위의 함수는 Chi\_Square\_Verification을 위한 값을 출력하는 함수다. 을 통해 에 해당하는 값을 반환하는 함수다.

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

위의 그림은 P\_Value에 해당하는 자유도 99 카이제곱 값을 나타낸다. 그리고 P\_Value\_Check 함수는 해당 값과 만든 변수의 카이값을 비교하여 출력해준다.

1. result

텍스트이(가) 표시된 사진

자동 생성된 설명

프로그램 실행 시 평균과 분산을 입력하도록 한다.

텍스트, 스크린샷, 전자기기, 컴퓨터이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 전자기기, 컴퓨터이(가) 표시된 사진

자동 생성된 설명

평균 0, 분산 1로 값을 입력했을 때 다음과 같이 값과 히스토그램이 출력되고 chi값을 출력하며 이를 P Value와 비교해준다. 위의 그림은 정규분포와 일치하는 비율이 조금 떨어지는 것을 알 수 있다.

텍스트, 스크린샷, 전자기기, 컴퓨터이(가) 표시된 사진

자동 생성된 설명

다음 그림은 동일하게 평균 0, 분산 1로 값을 입력하였을 때 P\_Value와 히스토그램을 통해 정규 분포에 더 가깝게 출력된 것을 볼 수 있다.

텍스트, 스크린샷, 전자기기, 컴퓨터이(가) 표시된 사진

자동 생성된 설명

다음 그림은 평균 100 분산 25를 입력하였을 때의 모습이다. 해당 값이 100을 기준으로 근처에서 생성된 것을 히스토그램을 통해 관찰할 수 있다. 그리고 범위는 분산이 커짐에 따라 커진 것을 볼 수 있다.

텍스트, 스크린샷, 전자기기, 컴퓨터이(가) 표시된 사진

자동 생성된 설명

해당 그림은 평균 5, 분산 0.01을 입력하였을 때의 모습이다 해당 그림 또한 평균에 밀집되어 있고, 분산이 작기 때문에 더욱 조밀하게 모여 있는 것을 볼 수 있다.

1. 고찰

해당 과제를 통해 CDF와 PDF의 관계에 대해 더욱 자세히 알 수 있는 기회였고, 표준 분포에 대해서도 자세하게 공부할 수 있었다. 그리고 난수를 만들 때에 정규 분포로 만들기 위한 다양한 방법이 존재한다는 점을 알게되었고 이러한 식이 어떻게 유도되는지를 관찰할 수 있었다. 그리고 카이제곱 검증을 통해 필자가 만든 값이 정확한지 검증하는 과정에서 수학적인 사고를 넓힐 수 있었다. 기본적으로 박스-밀러 변환은 N(μ,σ^2)을 따르기 위해서는 PI와 cos, sin 연산을 사용할 수 밖에 없다. 하지만 필자가 사용한 방법을 통해서는 이러한 연산을 줄이는 대신 루트 연산을 늘린다. 하지만 해당 방식에는 단점이 존재한다. 이는 해당 방식을 사용하였을 때 필요보다 많은 난수를 생성해야하는 것이다..

1. Reference

* Box-Muller transform :

<https://kplus-biz.github.io/%EC%98%88%EC%A0%9C/2017/10/29/Box-Muller-transformation/>

<https://en.wikipedia.org/wiki/Box%E2%80%93Muller_transform>

* Marsaglia polar method : <https://en.wikipedia.org/wiki/Marsaglia_polar_method>

<https://tibyte.kr/219>

* 카이제곱 분포 : <https://terms.naver.com/entry.naver?docId=3338174&cid=47324&categoryId=47324>

<https://namu.wiki/w/%EC%B9%B4%EC%9D%B4%EC%A0%9C%EA%B3%B1%EB%B6%84%ED%8F%AC?from=%EC%B9%B4%EC%9D%B4-%EC%A0%9C%EA%B3%B1%20%EA%B2%80%EC%A0%95>