○ 실험보고서

확률과통계 HW#2

제출일 : 06월 04일

2018008559 컴퓨터소프트웨어학부 신상윤

1. Introduction

1.1 실험목적

 두 변수 X, Y의 상관계수를 분석한다. 두 변수의 marginal, joint distribution(확률분포), x-y 2차원 평면상에서 두 변수의 분포 그래프를 그리고, 표준편차 σ를 3가지로 다르게 하여 상관계수를 각각 구하고 비교해본다. 이때 발생하는 변수 x 값의 범위와 정밀도, 표준편차의 크기 설정에 주의한다.

1.2 배경 이론

1.2.1 Poisson 분포

 푸아송분포는 단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 표현하는 이산 확률분포이다. 정해진 시간 안에 어떤 사건이 일어날 횟수에 대한 기댓값을 λ라고 했을 때, 그 사건이 n회 일어날 확률을

$P(k) = \frac{\lambda^k}{k!}e^{-\lambda}$로 정의한다. 이때 평균과 분산은 모두 λ이다.

이항 분포의 기댓값을 λ라 하자. $np = \lambda$, $p = \frac{\lambda}{n}$ 즉,

$P(X=x) = \binom{n}{x}(\frac{\lambda}{n})^x (1-\frac{\lambda}{n})^{n-x} = \frac{\lambda^k}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}(1-\frac{\lambda}{n})^{(-\frac{n}{\lambda})(-\lambda)}$에서

n이 ∞로 갈 때 $P(X=x) = \frac{\lambda^k}{k!}e^{-\lambda}$가 되고, 이는 푸아송분포와 같다.

1.2.2 normal 분포

  정규분포는 연속 확률분포 중 하나이며, 수집된 자료의 분포를 근사하는 데에 자주 사용된다. 이는 중심극한정리에 의하여 독립적인 확률변수들의 평균은 정규분포에 가까워지는 성질이 있기 때문이다. 확률 밀도함수는

$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ 평균과 분산은 각각 $E(X) = \mu$, $V(X) = \sigma^2$이다.

1.2.3 Box-Muller transform

C에서 따로 정규분포를 따르는 랜덤한 값을 제공해주지 않으므로 직접 코드로 구현해야 한다. C에서 제공해주는 것 중에서 하나가 rand()인데 이는 연속균등분포라 볼 수 있다. 연속균등분포에서 정규분포를 만들수있게 하는 것이 **Box Muller transform** 이다.

가장 핵심적인 내용은 U_1, U_2를 (0,1)에서 균등분포라 할 때 $Z_0 = \sqrt{-2\ln U_1}\cos(2\pi U_2)$,$Z_1 = \sqrt{-2\ln U_1}\sin(2\pi U_2)$는 각각 표준정규분포를 따른다는 것이다. 이를 이용해 정규분포를 구현할 수 있다.

1.2.4 Covariance and Correlation Coefficient

공분산 Cov(X, Y)는 다음과 같이 정의한다.

$$Cov(X, Y) = \sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$$

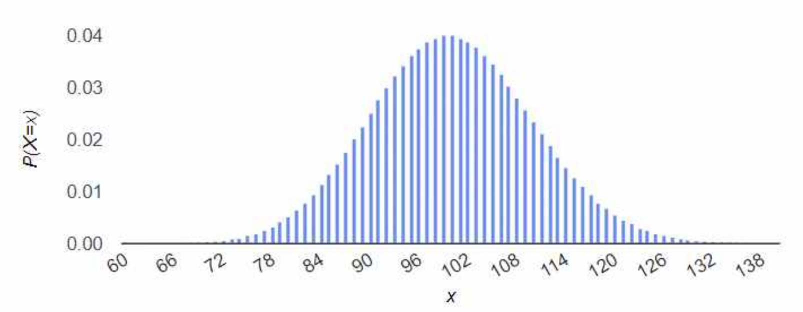
상관계수 ρ_{XY}는 다음과 같이 정의한다.

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

2. Process

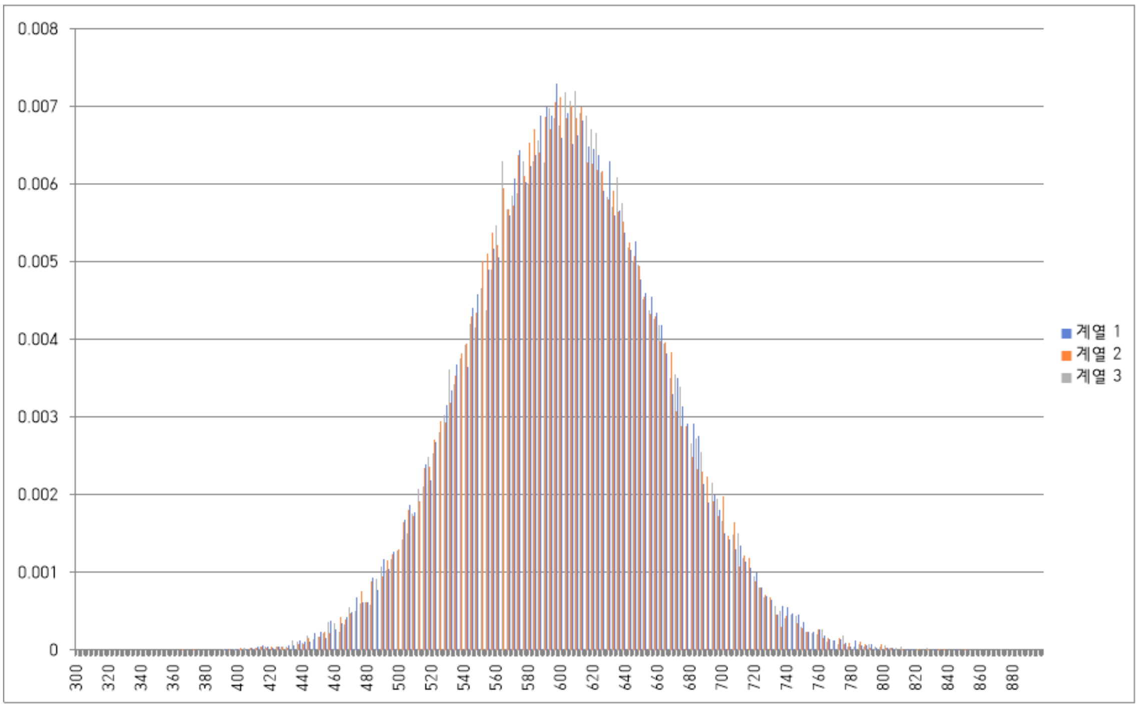
변수 X는 저번 실험에서 사용하였던 λ가 100인 푸아송분포에서 생성했고,
변수 Y는 제출일이 6월 4일이라 $Y = 6X + 4 + N(0,\sigma^2)$로 설정했다.
σ는 각각 10, 20, 30일때 실험했다.

X의 marginal distribution

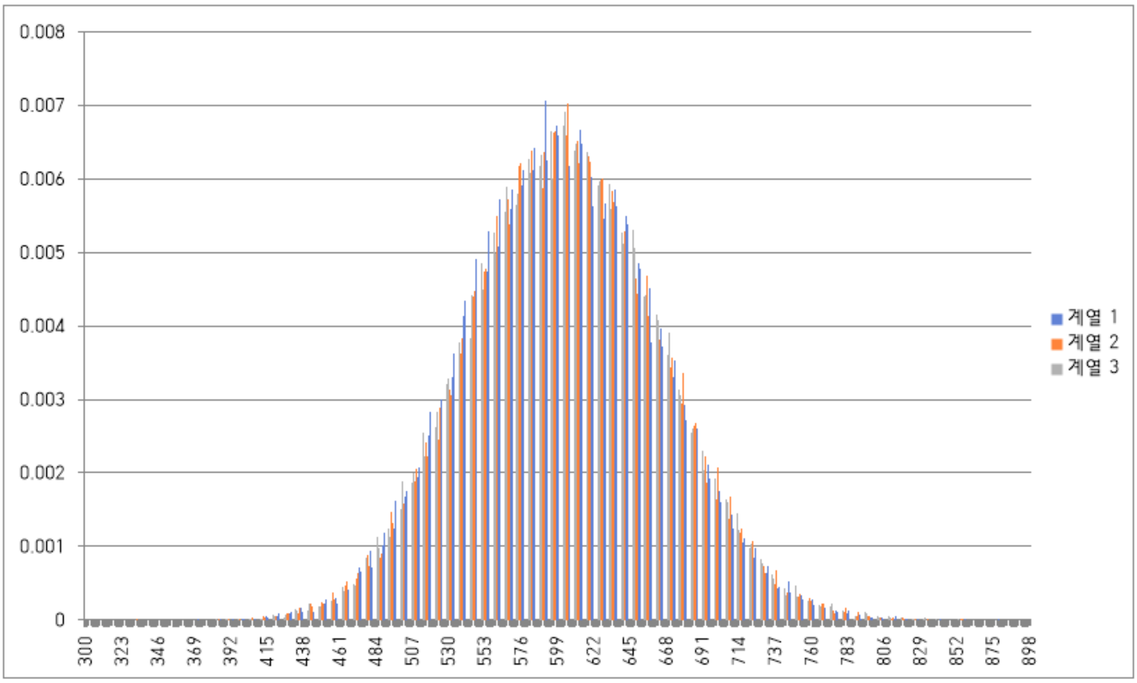


Y의 marginal distribution

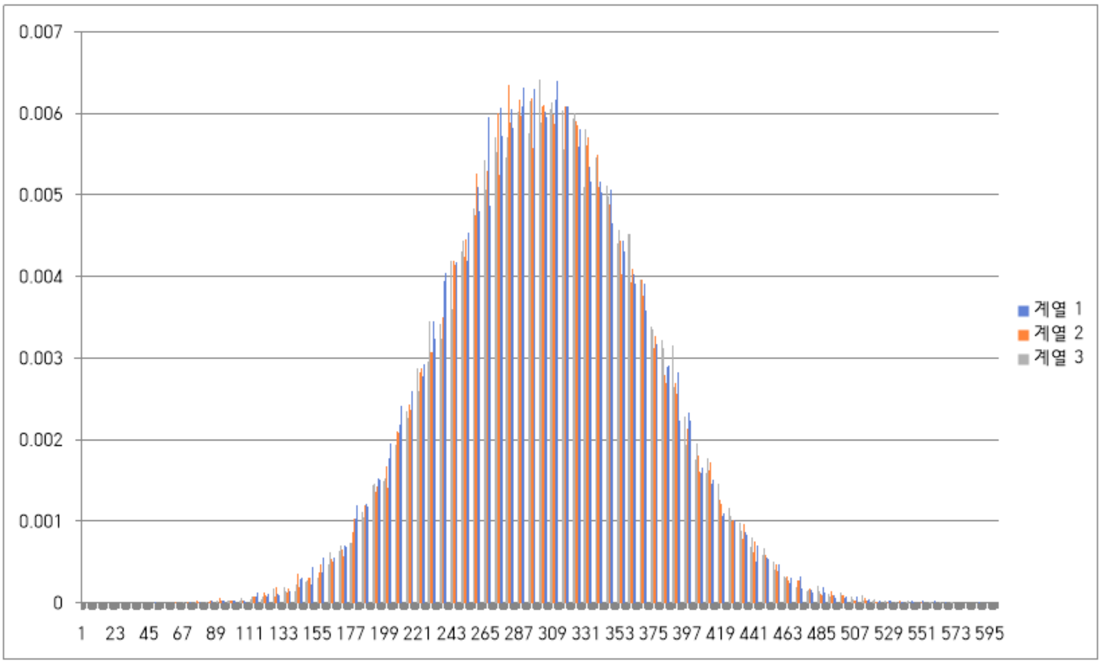
(σ = 10)



(σ = 20)

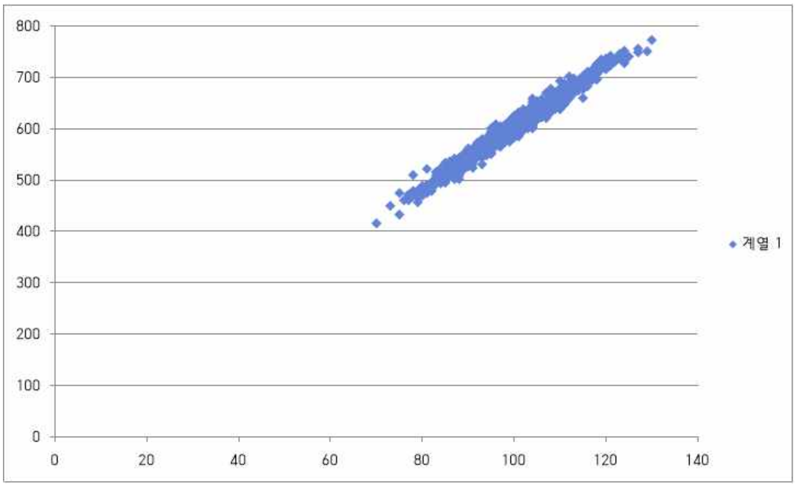


(σ = 30)

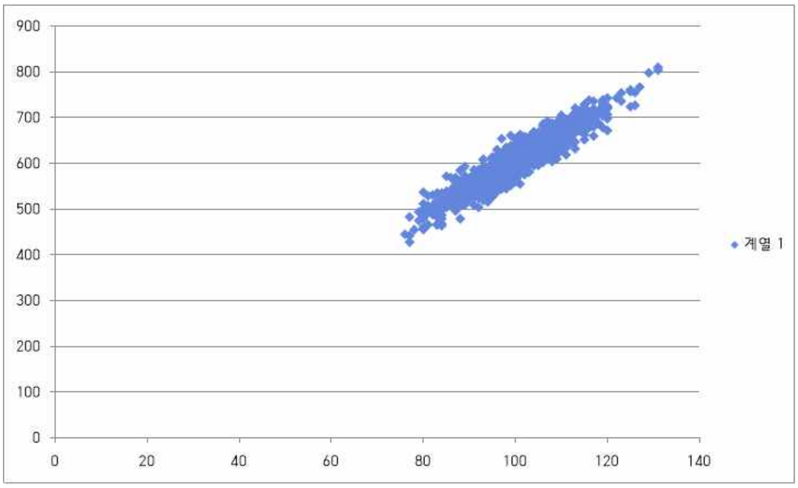


3. Result

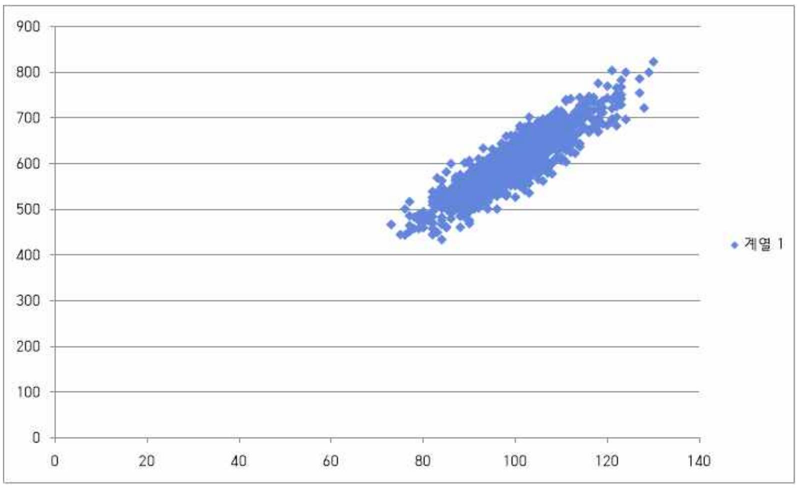
(σ = 10) ρ_{XY} = 0.9848



(σ = 20) ρ_{XY} = 0.9343



(σ = 30) ρ_{XY} = 0.8854

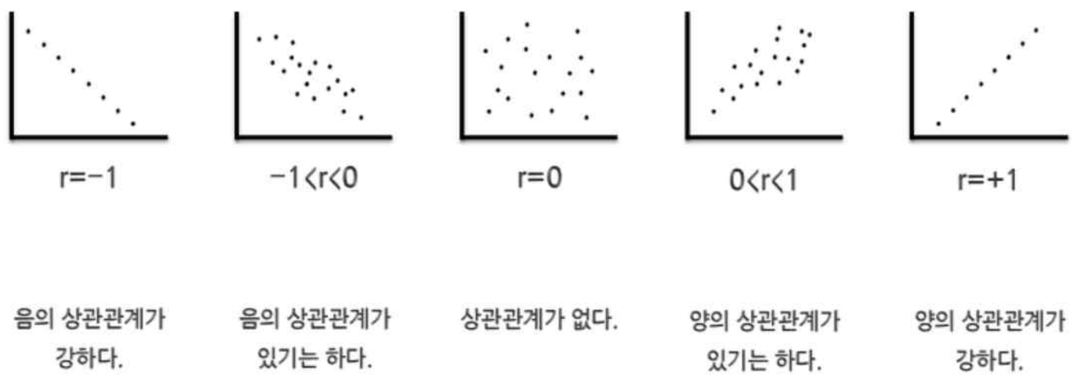


3.1 결과 분석

X의 분포를 푸아송분포로 잡았고, Y를 X의 상수배 + 상수 + 정규분포 꼴로 잡았다. 정규분포와 푸아송분포 모두 평균값이 나올 확률이 가장 높은 종 모양 분포를 이루고 있으므로 Y의 분포도 종 모양 분포가 나올 것이라고 예상할 수 있다. 실제로 실험을 통해 그린 Y의 그래프가 종 모양임을 확인할 수 있다.

σ가 각각 10, 20, 30일 때 그래프를 따로 구했는데 관찰해보면 σ가 커질수록 그래프가 낮아지고 퍼지는 것을 알 수 있다. 이는 정규분포에서 σ가 커질 때와 동일한 현상으로, 그려진 그래프에서 P(Y = 600)을 보면 σ가 10, 20, 30일 때 각각 0.007, 0.0065, 0.006으로 실제로 그래프가 낮아지는 것을 수치적으로 볼 수 있다.

이제 분포와 상관계수를 보면 σ가 커질수록 분포가 퍼지는 것을 볼 수 있다.



위 그림과 같이 우리가 구한 분포에서 σ가 커질수록 양의 상관관계가 작아짐을 알 수 있다. 상관계수는 절댓값이 1에 가까워질수록 상관관계가 강하다 할 수 있다. 실제 실험에서 각 분포에서 상관계수를 구해보면 σ가 커질수록 상관계수가 작아지는 것을 알 수 있다. 이는 σ가 커질수록 그만큼 정규분포에서 0에서 크게 벗어난 값이 나올 확률이 증가함으로 인하여 X, Y의 상관관계가 약해졌다고 생각할 수 있다. 또한 3가지 경우 모두 상관계수가 0.9 언저리로 높은 값인데 이는 Y가 정해지는 확률에 X가 영향을 끼치기 때문이라 예상할 수 있다. 만약 Y = 6X + 4라 하면 Y는 X에 의하여 무조건 결정되는 것이므로 상관관계가 1이라 할 수 있다. 즉, 상관관계는 오직 정규분포의 결과에 의해서만 결정된 것이므로 각 σ에 따라 상관계수, 분포 등이 달라지는 것임을 알 수 있다.

3.2 코드

```
#include <stdio.h>
#include <math.h>
#include <Windows.h>
#include <time.h>

double gaussianRandom(double average, double stdev){
  double v1, v2, s, temp;
  do {
    v1 =  2 * ((double) rand() / RAND_MAX) - 1;      // -1.0 ~ 1.0 까지의 값
    v2 =  2 * ((double) rand() / RAND_MAX) - 1;      // -1.0 ~ 1.0 까지의 값
    s = v1 * v1 + v2 * v2;
  } while (s >= 1 || s == 0);
  s = sqrt( (-2 * log(s)) / s );
  temp = v1 * s;
  temp = (stdev * temp) + average;
  return temp;
};

int main(void) {
    double a[1001] = {0};
    double b[1001] = {0};
    double s1=0,s2=0,s3=0;
    double mx,my;
    srand(time(NULL));
    for(int i=0; i<1000; i++){
        int cnt = 0;
        for(int j=0; j<1000; j++){
            if(rand()%10 == 0)
                cnt++;
        }
        b[i] = cnt;
        double n = 6*cnt + 4 + gaussianRandom(0, 10);
        //a[n]++;
        a[i] = n;
    }
    for(int i=0; i<1000; i++){
            s1 += a[i];
            s2 += b[i];
    }
    mx = s1/1000;
    my = s2/1000;
```

```
    s1 = 0;
    s2 = 0;
    for(int i=0; i<1000; i++){
        s1 += (a[i]-mx) * (b[i]-my);
        s2 += (a[i]-mx) * (a[i]-mx);
        s3 += (b[i]-my) * (b[i]-my);
    }
    float f = (s1*s1*1.0)/(s2*s3);
    printf("%.3f %.3f",s1,f);
    /*
    int cnt1 = 0;
    for(int i=300; i<900; i++){
        printf("%.5f\n",a[i]/100000);
        cnt1 += a[i];
    }
    printf("%d",cnt1);*/
    return 0;
}
```

4. Discussion

4.1 변수 X의 생성

위 실험에서는 X를 푸아송분포(λ = 100)의 확률변수라 잡았다. X를 만약 다른 분포들로 했으면 Y의 분포나 다른것들이 얼마나 달라질까? 사실 어떤 분포를 써도 중심극한정리에 의하여 결국 Y도 정규분포에 가까워지고 그 성질들을 보여줄 것이다. 상관계수 분석이라는 주제에서는 결과에 크게 영향을 미치지 않고 σ가 커질수록 똑같이 상관계수의 절댓값이 0에 가까워질 것이다.

4.2 a와 b

a와 b는 실험에 크게 영향을 끼치지 않는 상수들이다. 하지만 a와 b의 값에 따라서 그래프를 그리는데 범위가 커지고 값이 극단적일 수도 있게 되므로 a와 b는 1 ~ 9까지의 정수(혹은 -1 ~ -9)가 적당해 보인다.

4.3 σ의 값

실험을 통해 σ의 값에 따라 상관계수나 그래프 분포 등을 확인해보았다. σ의 차이가 클수록 차이가 명확히 보이는건 자명하다. 하지만 10 차이씩으로도 차이가 잘 보이는 것 같아 10, 20, 30도 적당해 보인다. 만약 다음 실험을 계획한다면 다음 실험에서는 σ를 10, 30, 50으로 하는 것이 좋아 보인다.

5. Reference

[1] https://ko.wikipedia.org/

[2] https://homepage.stat.uiowa.edu/~mbognar/applets/

[3] https://en.wikipedia.org/wiki/Box%E2%80%93Muller_transform