통계학 입문

2021-05-02

데이터과학융합스쿨

손 대 순 교수

이전시간에 몇가지 이산형 분포와 연속형 분포를 다루었습니다. 중심극한정리라는 것이 통계학에서 매우 중요한 정리임을 강조했었구요. 이번 시간에는 이항분포와 정규분포의 관계에 대해서 생각해 보고자 합니다.

앞서 이산형 확률분포를 이야기할 때, 포아송 분포를 설명했습니다. 이항분포에서 n이 크고 성공확률 p가 매우 작으면 q = (1-p) 값이 거의 1에 가까운 현상이 생기죠. 이 때 포아송 분포를 사용할 수 있다고 했습니다. n이 커지면 이항분포에서 확률계산하기가 어려워지는 문제가 생기기 때문이죠. n!의 계산이 너무 어려워집니다. 실제 그런지 한 번 해 볼까요?

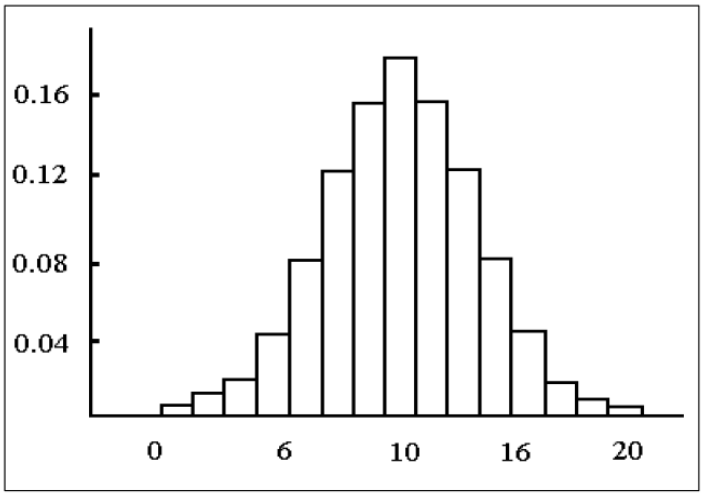
R에서 factorial을 계산하는 함수는 그대로 "factorial"입니다.

```
> factorial(10)
> factorial(100)
> factorial(1000)
```

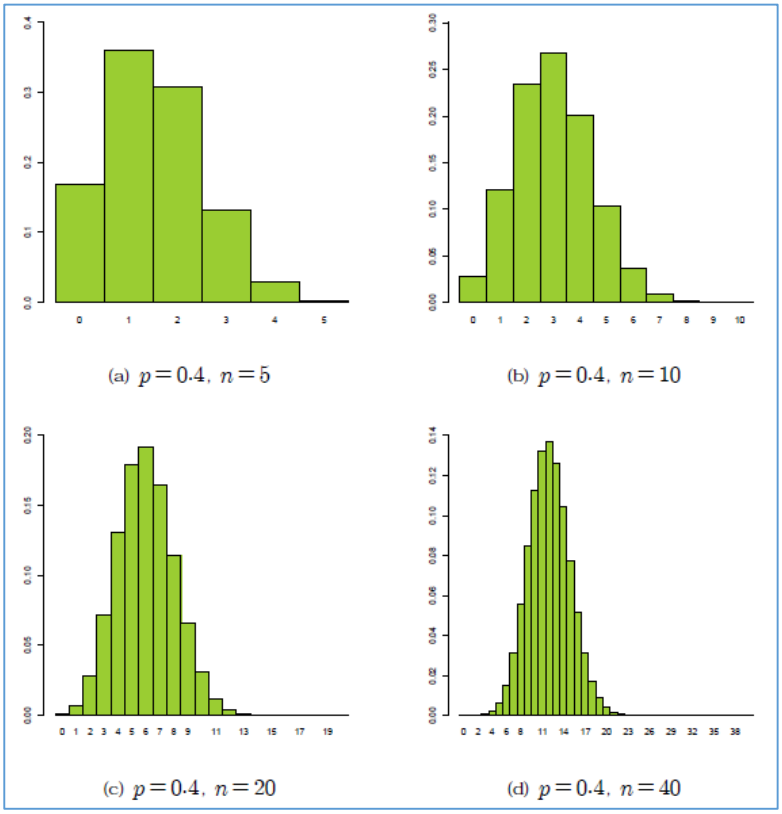
한 번 해 보세요. 어떤 일이 벌어지나요? 100!만 해도 어마어마한 숫자가 나올거예요. 읽을 수 있어요? ㅎㅎ 1000!은 더욱 계산 못할 겁니다. 무한대라는 표현을 쓰며 계산을 포기할거예요. 실제 그런지 여러분들이 한 번 해 보세요. 그러니, n이 커지면 이항분포의 확률밀도함수를 통해서 확률을 계산하는 것이 불가능해 집니다.

자, n이 큰 상황에서 p가 작으면 포아송 분포를 이용하여 확률 계산할 수 있다고 했습니다. 그런데, p가 작지 않다면 어떨까요? 그렇다면 확률 계산을 어떻게 할 수 있을까요? n이 크다면 여전히 factorial 계산이 어렵다는 문제를 가지고 있는데, p가 작지 않으니 포아송 분포를 사용할 수가 없게 됩니다. 어떻게 할 수 있을까요?

이항분포는 n이 커지면 정규분포와 비슷하게 되는 성질이 있습니다. 이산형 분포이기는 하지만, 연속형 분포로 근사시켜서 확률을 계산할 수 있어요. 무슨 이야기인지 자세히 살펴 봅시다.



이 그림은 B(20, 1/2)을 따르는 확률분포입니다. 어떤가요? 정규분포와 비슷한가요? p=1/2이어서 이렇게 되는건지도 한 번 확인해 봅시다.



p=0.4인 경우도 n이 커지니까 점점 정규분포와 비슷한 모양이 되어 가네요? 그렇죠? 이항분포는 n이 커지면 정규분포와 가까워 집니다. 그렇다면 정규분포를 이용해서 근사적으로 확률계산이 가능하겠네요? 맞습니다. 그런데, 한가지 고려해야할 부분이 있어요. 이항분포는 이산형 분포죠. 동전을 n번 던져 앞면이 x번 나올 확률 같은 것이니, 확률 변수 X가 가지는 값이 1,2,3,4와 같이 정수형입니다. 따라서, 이를 연속형 분포에 근사시켜 확률 계산을 할 때는 '연속성 수정(continuity correction)'이라는 작업을 해 줍니다. 어렵지 않아요.

X가 이항분포 $B(n,p)$를 따를 경우, $P(X=k)$(단, k=정수)를 $P(k-0.5<X<k+0.5)$으로 수정

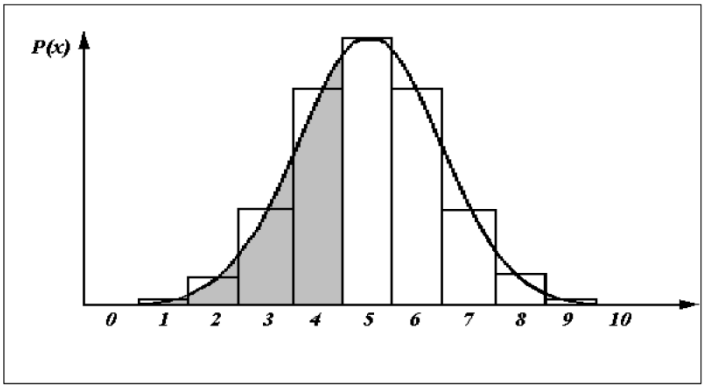
위에서 쓰여진대로, 0.5를 더하고 빼주어 범위의 형태로 만들어 준다는 말입니다. 즉, 이산형 분포에서

$$P(1 \le X \le 3) = P(X=1) + P(X=2) + P(X=3)$$

위와 같이 표현할 수 있겠지요? 이를 연속형 분포로 근사시켜 계산한다고 해 봅시다. 위와 같은 범위의 확률을 계산한다고 하면 확률에 오차가 조금 커질 수 있어요. 그래서 위 범위에서 앞뒤로 0.5만큼씩 더 계산해 주는 것입니다.

$$P(0.5 \le X \le 3.5)$$

위와 같이 확률계산의 범위를 보정해 주는 것을 연속성 수정이라고 합니다.



위의 그림은 B(10, 0.5)의 확률분포 히스토그램과 N(5, 2.5)의 확률분포를 함께 표시한 그래프예요. 히스토그램에서 확률계산을 할 때와 정규분포에서 확률계산할 때, 그래프 면적의 차이를 보세요 (회색 부분). 연속성 수정을 해 주면, 면적에서 빠진 부분이 보완되는 현상을 볼 수 있죠? 좀 더 정확한 값에 가까운 근사치를 얻어내기 위한 방법입니다. 개념적으로는 그리 어렵지 않죠?

n이 클수록 이항분포를 따르는 확률변수 X에 대한 확률 계산은 정규분포로 근사시켜 계산할 수 있다고 했습니다. 그렇다면 정규분포의 모수는 어떻게 알 수 있을까요? 아마 위의 그림으로 짐작한 사람들이 있겠네요.

$$X \sim B(n,p) \quad \approx \quad X \sim N(np, npq)$$

위와 같이 되겠지요. 이항분포의 평균이 np이고 분산이 npq이니까 말입니다. 간단한 문제를 하나 풀어 볼까요?

어느 양계장에서는 병아리를 부화시키는데 5% 정도의 실패율을 가지고 있다. 새로 500개의 달걀을 부화시키는데 실패율이 3% 이하일 가능성은 얼마인가?

부화에 실패한 달걀의 개수를 확률변수 X로 두면 되겠네요. 그렇다면 확률변수 X는 다음과 같은 이항분포를 따릅니다.

$$X \sim B(500, 0.05)$$

실패율이 3%이하일 확률을 구하라고 했군요. 500개의 3%이면 15개입니다. 그러니 부화 실패한 달걀의 개수가 15개 이하일 확률을 구하라는 말입니다.

$$P(X \leq 15) = \sum_{x=0}^{15} \binom{500}{x} (0.05)^x (0.95)^{n-x}$$

정확히 계산하기 위해서는 위와 같은 식을 이용해야겠지요. 그런데 앞서서 해 보았던 것처럼, 500!을 계산하기는 쉽지 않습니다. 자, 이럴 때 정규분포에 근사시켜서 확률을 계산할 수 있어요.

정규분포는 연속형 분포이니 연속성수정을 하면

$$P(X \leq 15.5)$$

이 확률을 구하면 되겠군요. 정규분포에 필요한 모수는 평균과 분산입니다. 이는 이항분포에서 금방 계산할 수 있지요.

$$E(X) = np = 500 \times 0.05 = 25$$

$$Var(X) = npq = 500 \times 0.05 \times 0.95 = 23.75$$

$$X \sim N(25, 23.75)$$

정규분포를 따른다는 것을 알고 모수를 확인하였으니, 이제 확률을 계산하면 됩니다. 정규분포에서는 표준화 과정을 통해 확률을 금방 찾아낼 수 있죠.

$$P(X \leq 15.5) = P\left(Z < \frac{15.5 - 25}{\sqrt{23.75}}\right) = P(Z < -1.95) = 0.0256$$

이는 표에서 찾을 수 있겠죠? 꼭 직접 해 보아야 합니다.

R을 활용해서 이항분포의 확률값과 정규분포 근사한 확률값을 한 번 비교해 볼까요?

확률변수 X가 B(150, 0.6)의 분포를 따른다고 합시다.

$$P(82 \leq X < 102)$$

위의 확률을 구해 봅시다. 이항분포에서 확률을 구할 때는 102가 포함되지 않음에 유의해야 합니다. 150!정도는 아마 계산할 수 있을 것 같으니 이항분포에서의 확률계산을 해 보죠. R에서 이항분포의 확률 pbinom 함수를 이용할 수 있습니다.

R을 이용해서 다음을 수행해 봅시다.

```
> n=150
> p=0.6
> pbinom(101,n,p) - pbinom(81,n,p)

> mu = n*p
> sigma=sqrt(n*p*(1-p))
> pnorm(101.5,mu,sigma) - pnorm(81.5,mu,sigma)
```

위의 과정을 이해할 수 있겠죠? 이항분포를 이용해서 확률 계산해 보고, 정규분포를 이용해서 확률을 계산해 보았어요. 어때요? 비슷한 값이 나오나요?

R에서의 계산은 생각보다 참 간단합니다. 여러분들이 많이 익숙해 지기를 바래요.

오늘은 이항분포의 정규근사에 대해 공부했습니다. 다음 시간부터는 추정과 검정을 시작합니다. 이제 드디어 모집단에 대한 이야기를 본격적으로 하겠군요.^^

오늘도 수고 많았습니다.