Chapter 2.4 The Geometric Distribution

Bernoulli 랜덤변수와 같은 상황을 가정하자. 이 실험은 p 의 확률로 성공, (1-p)의 확률로 실패한다. 실험이 처음으로 성공할 때까지 걸리는 회수를 X 라고 할 때, X 는 Geometric Random Variable 이라고 한다.

Definition 2.8
확률 p 가 주어졌을 때, geometric random variable X 의 값은 다음과 같다. (n = 1,2,3 …)

$$\Pr(X = n) = (1-p)^{n-1}p$$

Bernoulli 때와 마찬가지로 가능한 모든 n 의 값을 모두 합하면, geometric 랜덤 변수도 1 이 나와야 한다. 이는 다음과 같이 증명이 가능하다.

Proof)

$$\sum_{i=1}^{\infty}(1-p)^{i-1}p = p\sum_{i=1}^{\infty}(1-p)^{i-1} = p * \frac{1}{1-(1-p)} = 1$$

Geometric 랜덤 변수의 큰 특징 중 하나는 memoryless 하다는 것이다. 즉, n 번을 시도해서 처음으로 성공한다고 가정해보자. 현재 K 번 시도하였다면, 지금부터 실험을 진행하여 n-k 번째 성공할 확률은 이전 실험과는 관계가 없다. (독립적이다.)

이 성질은 다음과 같이 식으로 표현할 수 있다.

$$\Pr(X = n + k \mid X > k) = \Pr(X = n)$$

Proof)

$$\begin{aligned}\Pr(X = n + k \mid X > k) &= \frac{\Pr((X=n+k)\cap(X>k))}{\Pr(X>k)} \\ &= \frac{\Pr(X=n+k)}{\Pr(X>k)} \\ &= \frac{(1-p)^{n+k-1}p}{\sum_{i=k}^{\infty}(1-p)^{i}p} \\ &= \frac{(1-p)^{n+k-1}p}{(1-p)^{k}} \\ &= (1-p)^{n-1}p \\ &= \Pr(X = n)\end{aligned}$$

Geometric 랜덤변수의 기대값은 다음과 같이 구한다.

E[X] = $\sum_{i=1}^{\infty} i(1-p)^{n-1}p$
= $p\sum_{i=1}^{\infty} i(1-p)^{n-1}$

이 때, $S = \sum_{i=1}^{\infty} i(1-p)^{n-1}$ 로 정하고, $S-(1-p)S = pS = \sum_{i=0}^{\infty}(1-p)^{i} = \frac{1}{p}$
따라서, $S = \frac{1}{p^2}$ 이고, E[X] = $\frac{1}{p}$ 이다.

Chapter 2.4.1 Example: Coupon Collector Problem

Geometric 랜덤 변수 대표적인 문제인 쿠폰 콜렉터를 정리한다.

시리얼 제조회사는 총 N 가지의 쿠폰을 만들었다. 그리고 시리얼을 팔 때, N 개의 쿠폰 중 1 개를 랜덤으로 시리얼 박스에 넣었다. 모든 쿠폰을 수집하기 위해서는 평균적으로 몇 개의 시리얼 박스를 구매해야 하는지 알아보는 것이 문제이다.

변수 X 를 N 개의 쿠폰을 다 수집하는데 필요한 박스의 개수라고 정하자. 그리고 X_i 는 i-1 개의 쿠폰을 수집하였고, i 번째 새로운 쿠폰을 수집하는데 필요한 박스의 개수라고 정의하자. (종류는 상관하지 않아도 된다.)

그러면 i=1, 2, …, n 에 대하여 각 X_i 는 geometric 랜덤 변수이다.
(새로운 쿠폰을 구할 때까지, 계속 박스를 사야하므로…)

박스를 샀을 때, 새로운 쿠폰이 나올 확률 $p_i = 1 - \frac{i-1}{n}$ 이다.
(i-1 개의 쿠폰을 수집한 상태)

$$E[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$$

$$\mathrm{E[X]} = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n\sum_{i=1}^{n} \frac{1}{i} = n\ln n$$

($\sum_{i=1}^{n} \frac{1}{n-i+1} = \sum_{i=1}^{n} \frac{1}{i}$ 인 이유는, 좌변의 식은 우변의 식을 거꾸로 더한 것이기 때문이다.)

(* $\int_{x=1}^{n} \frac{1}{x} dx = \ln x$)

따라서, 총 구매해야하는 예상 박스의 개수는 $n\ln n$ 개 이다.