

확률과 통계

- ~~기하~~, 이항, 푸아송 분포 -

* 주의: 기하분포는 가우지 않음.

참고서: Head First Statistics

(7장 ~~기하~~, 이항, 푸아송 분포)

작성자 : 이계식

환경대학고, 컴퓨터공학과



이항분포

(p. 333, 336, 337)

동일한 시도를 n번 시도할 때, 특정 사건이

n번 나올 확률들의 분포를 이항분포라 한다.

조건: 각각의 시도는 상호 독립적이며, 각각의 시도에서
특정 사건이 발생할 확률은 동일하다.

예제: 4자선다 문제

4자리 질문을 모두 암으로 짜았을 때 맞출 확률은 $P = \frac{1}{4}$

질문:

n개의 질문을 모두 암으로 짜어서 맞출 때 n개의 정답을 얻을 확률?

맞히는 개수 $\rightarrow r$ | 0 | 1 | ... | r | ... | ... | $n-1$ | n

맞히는 개수 $\rightarrow r$								
$P(X=r)$								

r 개 맞힐 확률 \rightarrow

$$\underbrace{P(X=r)}$$

어떻게 계산하나?

$P(X=r)$ 계산 예제 : $n=3$ 인 경우

즉, 3 개의 문제 중에 r 개를 맞힐

확률.

여기서 $r \in \{0, 1, 2, 3\}$ 임.

따라서 4 가지 경우를 살펴보아야 함.

① 모두 틀린 경우

$$P(X=0) = \left(\frac{3}{4}\right)^3 = {}^3C_0 \cdot \left(\frac{1}{4}\right)^0 \cdot \left(\frac{3}{4}\right)^3$$

② 한 문제 맞힐 경우

$$P(X=1) = 3 \cdot \frac{1}{4} \cdot \underbrace{\left(\frac{3}{4}\right)^2}_{\substack{\text{3 경우 각각 발생할} \\ \text{ 확률}}} \quad \begin{array}{l} \text{3 경우 각각 발생할} \\ \text{ 확률} \end{array}$$

$\underbrace{{}^3C_1}_{\substack{\text{3 문제 중에 1개 선택.}}}$

③ 두 문제 맞힐 확률

$$P(X=2) = \underbrace{{}^3C_2}_{\substack{\text{3 문제 중에 2개 선택}}} \cdot \underbrace{\left(\frac{1}{4}\right)^2}_{\substack{\text{3 경우 각각} \\ \text{ 발생할 확률}}} \cdot \underbrace{\frac{3}{4}}_{\substack{\text{3 경우 각각} \\ \text{ 발생할 확률}}}$$

④ 세 문제 모두 맞힐 확률

$$P(X=3) = \left(\frac{1}{4}\right)^3 = {}^3C_3 \cdot \left(\frac{1}{4}\right)^3 \cdot \left(\frac{3}{4}\right)^0$$

$$\Rightarrow P(X=r) = {}^3C_r \cdot \left(\frac{1}{4}\right)^r \cdot \left(\frac{3}{4}\right)^{3-r}$$

일반화

한 번의 시도에서 성공할 확률 = P

위와는 결과가 발생한다는 의미

n 번 시도에서 r 번 성공할 확률

nC_r 경우의 수
발생할 확률

$$P(X=r) = {}^nC_r \cdot P^r \cdot (1-P)^{n-r}$$

$$= {}^nC_r \cdot P^r \cdot q^{n-r} \quad (q = 1 - P)$$

n 개 중에서
 r 개를 선택하는
방법의 수

선택한 r 개가
성공할 확률

r 개를 제외한
나머지가
실패할 확률

이런 경우 X 는 이항분포를 따른다 말하고

아래와 같이 표기한다.

$$X \sim B(n, P)$$

이항분포

만족 횟수

한 번의 시도에서
성공할 확률

이항 분포 적용 예제 (교재에 있는)

하나의 정상적인 동전을 30번 던져서 24번

앞면이 나온 확률은?

전통방법:

정상적인 동전을 30번 던져서 6번면이 나오는
횟수에 대한 확률은 X 는 이항분포를
따르며, 아래 관계를 만족시킨다.

$$X \sim B(30, \frac{1}{2})$$

따라서 앞면이 24번 나온 확률은 다음과
같다.

$$\begin{aligned} P(X=24) &= {}^{30}C_{24} \cdot \left(\frac{1}{2}\right)^{24} \cdot \left(\frac{1}{2}\right)^6 = {}^{30}C_{24} \cdot \left(\frac{1}{2}\right)^{30} \\ &\approx 0.0006 \end{aligned}$$

즉, 10,000회의 6의 확률이다.

또한 앞면이 24회 이상 나온 확률은 다음과 같다.

$$\begin{aligned} P(X \geq 24) &= P(X=24) + \dots + P(X=30) \\ &= 0.0007 \end{aligned}$$

즉, 10,000회의 7 수준이다.

{ 일반적으로 사건 발생 확률이 5% 이하면 그런 사건이
실제로 발생하기 어렵다고 본다. 따라서 동전을 30번 던져서
앞면이 24번 나온다면 그 동전은 정상적이지
않다고 판단한다.

결론
↓

이후에 가설검증을

배울 때 자세히 다룬다.

이항분포의 기대치와 분산 (p.338 ~ 341)

$X \sim B(n, p)$ 라 하자.

그러면 :

$$E(X) = n \cdot p$$

$$\text{Var}(X) = n \cdot p \cdot q \quad (q = 1 - p)$$

=====

설명 :

① X_1 을 한 번 시도한 경우의 확률분포라 하자.

즉,

r	0	1
$P(X=r)$	q	p

한 번 시도하면
성공 또는 실패만
발생함.

$$\text{따라서}, E(X_1) = 0 \cdot q + 1 \cdot p = p$$

$$\begin{aligned}
 \text{Var}(X_1) &= (0-p)^2 \cdot q + (1-p)^2 \cdot p \\
 &= p^2 \cdot (1-p) + (1-2p+p^2) \cdot p \\
 &= p^2 - p^3 + p - 2p^2 + p^3 \\
 &= p - p^2 = p(1-p) \\
 &= p \cdot q
 \end{aligned}$$

② X 와 X_1 , 사이에 아래 관계가 성립한다.

X 는 X_1 을 반복하여 독립적으로 n 번 실행하는 확률분포,
즉, X_1 을 n 번 독립관측 하는 것이다.

따라서, 아래의 식을 얻는다.

$$X = \underbrace{X_1 + \dots + X_1}_{n \text{번}}$$

③ 위 식을 이용하여 $E(X)$ 와 $\text{Var}(X)$ 를 아래와 같이
계산할 수 있다.

$$E(X) = E(\underbrace{X_1 + \dots + X_1}_{n \text{번}})$$

$$= n \cdot E(X_1)$$

$$= n \cdot p$$

$$\text{Var}(X) = \text{Var}(X_1 + \dots + X_n)$$

$$= n \cdot \text{Var}(X_1)$$

$$= n \cdot p \cdot q$$

④

이항 분포 예제

① 동전을 1번 던져 앞면이 1번 나올 확률들의

분포:

$$X_{coin} \sim B(n, \frac{1}{2})$$

② 주사위를 1번 던져 2보다 큰 소수가 1번
나올 확률들의 분포:

$$X_{dice} \sim B(n, \frac{1}{3})$$

이때: 2보다 큰 소수는 3과 5이다.

포아송 분포

(P. 346 ~ 348)

특정 기간 동안 어떤 사건이 발생하는 횟수의 평균값 또는 비율이 주어졌을 때, 미래의 동일한 기간 동안 해당 사건이 r번 발생할 확률들의 분포

예제) 영화관의 낡은 펑크기계

- 주중에 평균 3.4회 고장 발생
 $\xrightarrow{\text{특정 기간}}$ $\xrightarrow{\text{사건 발생 번도수의 평균값}}$

질문: 다음 주중에 한 번도 고장이
발생하지 않을 확률은?

해답: 이런 경우 포아송 분포를 활용한다.

펑크기계의 경우 아래처럼 표기한다.

$$X \sim P_0(3.4)$$

일반화

어떤 주어진 구간에 사건이 발생하는 수를

X 라 할 때, 만약 X 가 구간마다 λ 번

발생하는 평균값을 갖는다면,

→ 같다라 하겠다

X 를 푸아송 분포를 따른다고 말하고

아래와 같이 표기한다.

그리스어 알파벳
(영어의 L에 해당)

$$X \sim P_0(\lambda)$$

X 가 푸아송 분포를 따를 때 $P(X=r)$ 은
아래와 같다.

$$P(X=r) = \frac{e^{-\lambda} \cdot \lambda^r}{r!}$$

($e \approx 2.71$, 오일러 상수)

특정 구간에 있음의
사건이 r 번 발생할 확률

예제 9: 영화관 흥행기록 (P.349~350)

영화관 흥행기록의 고장률은 평균 1회를

따른다: $X \sim P_0(3.4)$

여기서, 다음 주에 흥행기록이 한 번도 고장나지
않을 확률은 다음과 같다.

$$\begin{aligned} P(X=0) &= \frac{e^{-3.4} \cdot (3.4)^0}{0!} \\ &= \frac{1}{e^{3.4}} \\ &\approx 0.033 \end{aligned}$$

즉, 3% 정도이다.

결론: 아래와 세 흥행기록을 조사하는 데
좋을 듯 같다.

주어진 분포의 기대치와 표준偏差 (P. 34f)

주어진 분포의 기대치와 표준偏差은 매우 관련하다.

$$X \sim P_0(\lambda) \text{ 인 경우}$$

$$\begin{aligned} E(X) &= \lambda \\ \text{Var}(X) &= \lambda \end{aligned}$$

이제,

상식적으로 생각할 후
있는 경우와 일치한다.

경험적으로 특정기간에
평균 2번 어떤 사건이
발생하였다면, 미래에도
동일한 기간동안 2번
동일한 사건이 발생하는 것이라
기대할 수 있다.

푸아송 분포의 결합 (P. 352 ~ 353)

X 와 Y 가 독립 확률 변수일 때 아래가 성립한다.
(5장 참조)

$$E(X+Y) = E(X) + E(Y)$$

따라서 두 개의 독립 확률 변수 X, Y 에 대해서도
아래 관계가 성립한다.

$$\begin{aligned} X &\sim P_0(\lambda_x) \\ Y &\sim P_0(\lambda_y) \end{aligned}$$

두 사건이 서로 독립이며
동일한 가능을
대상으로 해야 함.

$$\Rightarrow \begin{cases} X+Y \sim P_0(\lambda_x + \lambda_y) \\ E(X+Y) = \text{Var}(X+Y) = \lambda_x + \lambda_y \end{cases}$$

예제 : 광온가지

$$X \sim P_0(3.4)$$

광온가지
일주연동안
보강횟수

음온가지

$$Y \sim P_0(2.3)$$

온온가지
일주연동안
고장횟수

서로 독립임.

두 가지 중 어느 하나의 일주연동안의 고장횟수는
 $X+Y \sim P_0(3.4 + 2.3)$ 의 푸아송 분포를 따른다.

따라서, 예를 들어, 다음의 일주연동안 어느 기계도
고장나지 않고 일주연동안은 아래와 같다.

$$P(X+Y=0) = \frac{e^{-5.7} \cdot 5.7^0}{0!} = 0.003 \quad (0.3\%)$$

이항분포와 농아승 분포의 관계 (P.356)

$X \sim B(n, p)$ 이고 $n \geq 50$, $p \leq 0.1$ 인

경우 $X \sim P_0(np)$ 라고 해도 무방하다.

즉, 이항분포인 경우 개별 적당히 크고, P 가

매 적으면 농아승 분포를 이용하여

확률을 계산할 수 있다.

예제 (P.357)

50개의 원자를 무작정 찍어서 5개의 양을
갖춘 확률은 얼마인가?

만, 각 원자를 찍어 정답을 갖춘 확률은
0.05이다.

질문 X 가 정답 개수라 할 때 $X \sim B(50, 0.05)$
이다. 따라서, 농아승 분포를 이용할 수 있다.

즉, $X \sim P_0(2.5)$ 라고 보자.

$$\text{따라서, } P(X=5) = \frac{e^{-2.5} \cdot (2.5)^5}{5!} = 0.067$$

즉, 6.7%이다.

□