3.7 초기하 분포(Hyper-geometric)

정의(DEFINITION)

크기 N 모집단은 서로 다른 두 집단으로 구성되어 있고 한 그룹은 M 개 원소, 다른 그룹은 $(N-M)$ 개 원소로 구성되어 있다. 모집단으로부터 크기 $n(\le N)$ 의 표본을 뽑을 때 M 개 원소인 그룹에서 뽑힌 원소의 개수를 확률변수 X 라 정의할 때 확률밀도함수는 다음과 같고 이를 초기하 분포라 한다.

$$f(x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},\ x=0,1,\dots,n,\ x\le M,\ n-x\le N-M$$

Notation $X \sim HG(N,M,n)$

IN SAS, PDF('HYPER',x,N,M,n);

정리(THEOREM)

(1)만약 $X \sim HG(N,M,n)$ 이면, $\mu=E(X)=n\frac{M}{N}$, $\sigma^2=V(X)=n\frac{M}{N}\frac{(N-M)}{N}\frac{(N-n)}{(N-1)}$.

(2)As $n\to\infty$, $HG(N,M,n)\sim(app)Binomial(n,\frac{M}{N})$ (증명 skipped)

EXAMPLE 3-13 초기하 분포 이용

20명의 학생 중 10명을 무작위(랜덤)로 뽑아 면접 조사를 실시한다. 10명을 뽑았을 때 20명 중 가장 뛰어난 5명이 포함되어 있을 확률을 계산하시오.

0.01625

EXAMPLE 3-14 초기하 분포 이용(2)

사과 한 상자에 20개가 들어 있다. 5개를 임의로 골라 상태를 보고 불량 사과가 없으면 구입 한다. 만약 4개의 불량 사과가 있는 상자를 골라 검사할 때

(1)사과 상자를 구입하지 않을 확률은?

(2)검사하는 5개 사과 중 불량 사과의 기대 개수는?

0.2487 / 1

HOMEWORK #7-2 DUE 4월 7일(목)

OO라디오에는 6개의 트랜지스터가 있고 그 중 2개는 불량이라고 하자. 3개의 트랜지스터를 임의 선택하여 검사할 때 불량의 개수를 확률변수 X 라 정의할 때 확률밀도함수를 얻으시오.

HOMEWORK #7-3 DUE 4월 7일(목)

초원에 서식하는 얼룩말 수(N)를 추정하려고 한다. M마리 얼룩말을 잡아 표식을 붙이고 놓아 주었다. 일정 기간이 지난 후 n마리 얼룩말을 잡아 표식 여부를 확인하였다. n마리 중 표식이 있는 얼룩말의 수를 확률변수 X 라 정의하자. 그리고 M=4, n=3이라 가정하자.

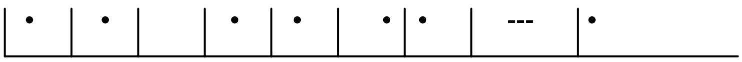
(1) $P(X=1)$ 일 확률을 구하시오.

(2) $P(X=1)$ 을 최대화 하는 모집단 얼룩말 수 N을 추정하시오.

3.8 포아송 분포 (Poisson)

단위 시간, 면적에서 임의의 사건 성공 회수에 관심을 갖는 경우를 생각하자. 한남대 앞 정류장에 도착하는 버스 수(시간 당), 한 페이지 당 오타 숫자, 은행 창구를 찾는 고객 수(10분 당) 등이 예이다.

시간이나 면적을 각 구간에서는 많아야 하나의 사건이 있어 나도록 동일 크기의 구간으로 나누자. 구간의 수를 n 이라 하고 각 구간에서 사건이 발생할 확률을 p 라고 하자. 그러므로 각 구간에서 사건이 일어나지 않을 확률은 $(1-p)$ 이고 각 구간은 마치 베르누이 시행과 같은 실험이 된다.



$e^{-\lambda} = \lim_{n\to\infty}(1-\frac{\lambda}{n})^n$ 사실을 이용하여 다음을 증명할 수 있다. 모수 (n, p) 인 이항분포는 모수 $\lambda = np$ 인 포아송 분포에 근사한다.

n이 충분히 크고 p가 매우 작을 때 이항분포는 포아송 분포에 근사한다.

$$\lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x} = \lim_{n\to\infty}\binom{n}{x}(\frac{\lambda}{n})^x(1-\frac{\lambda}{n})^{n-x}$$

$$= \frac{\lambda^x}{x!}\lim_{n\to\infty}(1-\frac{\lambda}{n})^n \frac{n(n-1)...(n-x+1)}{n^x}(1-\frac{\lambda}{n})^{-x}$$

$$= \frac{\lambda^x}{x!}e^{-\lambda}$$

위의 확률밀도함수를 포아송 분포라 한다.

$$f(x)=\frac{\lambda^x}{x!}e^{-\lambda},\; x=0,1,2,\ldots$$

Notation $X \sim Poisson(\lambda)$

IN SAS, PDF('Poisson', x,λ);

정리(THEOREM)

$X \sim Poisson(\lambda)$ 일 때 $\mu=E(X)=\lambda,\ \sigma^2=V(X)=\lambda$ 임을 증명하시오. (포아송 분포는 평균과 분산이 동일한 분포이다)

PROOF

$$E(X)=\sum_{x=0} x\frac{\lambda^x e^{-\lambda}}{x!}=\lambda\sum_{x=1}\frac{\lambda^{x-1}e^{-\lambda}}{(x-1)!}$$
$$=\lambda\sum_{z=0}\frac{\lambda^z e^{-\lambda}}{z!}=\lambda(\because e^{\lambda}=\sum_{x=0}^{\infty}\lambda^x/x!)$$

분산에 대한 증명은 숙제로 남겨 둔다. Q.E.D

정리(THEOREM)

확률변수 $X_i\ \underset{\sim}{iid}\ Poisson(\lambda), i=1,2,\ldots,n$ 일 경우 $\sum_{i=1}^{n}X_i \sim Poisson(n\lambda)$ 이다. 이를 분포의 가법성(additivity)이라 한다. (증명 later)

EXAMPLE 3-15 포아송 분포 가법성

초원에 서식하는 얼룩말의 수는 1 Acre 당 평균 5마리이고 포아송 분포를 따른다고 하자. 10 Acre를 무작위 조사하였을 때 얼룩말을 하나도 보지 못할 확률은?

$1.9\text{x}10^{-22}$

EXAMPLE 3-16 포아송 분포와 이항분포 관계

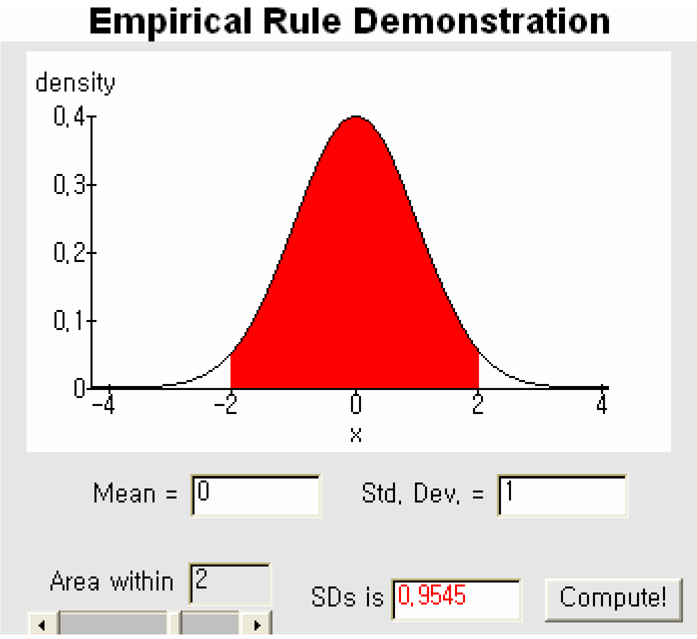
$X \sim Binomial(n=20, p=0.1)$ 일 때 $P(X\le 3)$ 을 구하는 문제에서
(1)이항분포를 이용하여 계산하시오 (2)포아송분포를 이용하여 계산하시오.

0.867 / 0.857

EXAMPLE 3-17 Empirical Rule

한남대 정문 앞 건널목에서 한 달 평균 3건의 사고가 발생하고 포아송 분포를 따른다고 가정하자. 그런데 지난 달에 6건의 교통사고가 발생하였다면, 교통사고의 평균 건수가 증가하였다고 할 수 있나? 아니다.

$6 \approx \mu + 2\sigma$ =6.46



$P(|X-\mu|<\sigma) = 0.68$
$P(|X-\mu|<2\sigma) = 0.95$
$P(|X-\mu|<3\sigma) = 0.99$

Tchebysheff's (보다 강화)

$$P(|X-\mu|<k\sigma) \ge 1-\frac{1}{k^2}$$

http://www.stat.sc.edu/~west/applets/empiricalrule.html

EXAMPLE 3-18 포아송 분포 이용

하루 8시간 작업 시간 중 기계가 멈춰 서는 회수는 평균이 2인 포아송 분포를 따른다.
(1)4시간 작업하였으나 기계가 멈추지 않을 확률을 계산하시오.
(2)하루 수익은 $(50-2X-X^2)$이다. 기대 수익을 계산하시오.

0.367 / 40

HOMEWORK #7-4 DUE 4월 7일(목)

$E(X(X-1))$을 이용하여 포아송 분포의 분산이 $V(X)=\sigma^2=\lambda$임을 증명하시오.

HOMEWORK #7-5 DUE 4월 7일(목)

$X \sim Poisson$ 이고 $P(X=1)=P(X=2)$ 이라면 확률 $P(X=4)$ 을 계산하시오.

HOMEWORK #7-6 DUE 4월 7일(목)

박물관을 방문하는 관람객 수는 평균 $\lambda = 2$(시간 당)인 포아송 분포를 따른다고 하자.

(a)오전 9:00~9:30 사이에 정확하게 3명의 관람객이 박물관을 찾을 확률은?

(b)오전 9:00~10:30 사이에 관람객이 한 명도 오지 않을 확률을 계산하시오.

(c)하루(8시간)에 몇 명의 관객이 오겠는가?

3.9 적률생성함수 (Moment Generating function)

정의(DEFINITION)

- 확률변수 X 에 대해 원점에 대한 k-차 적률(k-th moment)은 $E(X^k)$ 이다.
- 확률변수 X 에 대해 평균 μ 에 대한 k-차 적률은 $E((X-\mu)^k)$ 이다.
- 확률변수 X 의 적률생성함수, $M_X(t)$ 는 $E(e^{tX})$ 라 정의한다.

정리 THEOREM (왜 적률생성함수라 부르는가?)

$$M_x^{(k)}(t=0) = E(X^k)$$

PROOF

Taylor Series: $f(x) = f(0) + f'(0)x + \frac{f''(0)x^2}{2!} + \frac{f'''(0)x^3}{3!} + \ldots$

Taylor Series에 의해 $M_X(t) = E(e^{tx}) = E(1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \ldots)$ 이다.

그러므로 $M_x^{(k)}(t=0) = E(X^k)$ 임이 증명된다. Q.E.D.

Uniqueness of MGF (적률생성함수의 유일성)

동일 적률생성함수를 갖는 두 확률변수 X, Y 는 동일한 확률분포를 갖는다.

EXAMPLE 3-19 포아송 분포의 적률생성함수

확률변수 $X \sim Poisson(\lambda)$ 의 적률생성함수가 $e^{\lambda(e^t-1)}$ 임을 보이시오. 이를 이용하여 평균과 분산이 각각 λ, λ 임을 보이시오.

$$M_X(t) = E(e^{tX}) = \sum_x e^{tx}\frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda}\sum_x \frac{(\lambda e^t)^x}{x!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}$$

$$E(X) = M'(t=0) = e^{\lambda(e^t-1)}\lambda e^t |_{t=0} = \lambda$$

$$E(X^2) = M''(t=0) = e^{\lambda(e^t-1)}\lambda e^t \lambda e^t + e^{\lambda(e^t-1)}\lambda e^t |_{t=0} = \lambda^2 + \lambda$$

$$V(X) = \lambda$$

EXAMPLE 3-20 선형 함수의 적률생성함수

확률변수 X 의 적률생성함수가 $M_X(t)$ 일 경우 $(aX+b)$의 적률생성함수를 얻으시오.

$$M_{aX+b}(t) = e^{bt}M_X(at)$$

EXAMPLE 3-21 베르누이와 이항분포

$Bernoulli(p)$ 으로부터의 확률표본 $(X_1, X_2, ..., X_n)$ 의 합($\sum_{i=1}^n X_i$)의 분포가 $Binomial(n,p)$ 임을 보이시오.

$X \sim Bernoulli(p)$ 일 경우 $M_X(t) = E(e^{tx}) = \sum_x e^{tx}p(x) = q + pe^t$

$$M_{\sum X_i}(t) = E(e^{t\sum X_i}) = E(\prod e^{tX_i}) = (independent)\prod E(e^{tX_i}) = \prod(pe^t + q) = (pe^t + q)^n$$

이항분포의 적률생성함수와 같으므로 적률생성함수의 유일성에 의해 $\sum_{i=1}^n X_i$ 는 이항분포

EXAMPLE 3-22 포아송 분포 가법성

$Poisson(\lambda)$ 으로부터의 확률표본 $(X_1, X_2, ..., X_n)$ 합($\sum_{i=1}^n X_i$)의 분포는 $Poisson(n\lambda)$ 이다.

$$M_{\sum X_i}(t) = E(e^{t\sum X_i}) = E(\prod e^{tX_i}) = (independent)\prod E(e^{tX_i}) = \prod e^{\lambda(e^t-1)} = e^{(n\lambda)(e^t-1)}$$

HOMEWORK #8-1 DUE 4월 13일(화)

(a)확률변수 $X \sim Binomial(n, p)$ 의 적률생성함수가 $(pe^t + q)^n$ 임을 보이시오. 적률생성함수를 이용하여 평균과 분산인 각각 np, npq 임을 보이시오.

(b)확률변수 $X \sim Geometric(p)$ 의 적률생성함수는 $\frac{pe^t}{1-qe^t}$ 임을 보이시오. 적률생성함수를 이용하여 평균이 $/ p$ 임을 보이시오.

(c) $Geometric(p)$ 으로부터의 확률표본 $(X_1, X_2, ..., X_n)$ 의 합($\sum_{i=1}^{n} X_i$)의 분포가 $NB(n, p)$ 임을 보이시오. 그러므로 음이항분포 $NB(n, p)$ 의 적률생성함수는 $(\frac{pe^t}{1-qe^t})^n$ 이다.

HOMEWORK #8-2 DUE 4월 13일(화)

아래 적률생성함수를 갖는 확률분포함수를 얻으시오. (확률분포함수 참고하시오)

(a) $M(t) = [(1/3)e^t + (2/3)]^5$. (b) $M(t) = \frac{e^t}{2-e^t}$ (c) $M(t) = e^{2(e^t - 1)}$

3.10 Tchebysheff Inequality (체비세프 부등식)

정리(THEOREM)

확률변수 X 는 평균 μ , 분산 σ^2 갖는다고 하자. 양의 상수 k 에 대해 다음이 성립한다.

$$P(|X - \mu| < k\sigma) \ge 1 - \frac{1}{k^2} \quad or \quad P(|X - \mu| \ge k\sigma) \le \frac{1}{k^2}$$

PROOF

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\
&= \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)dx \\
&\ge \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x)dx + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(x)dx \\
&\quad \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 f(x)dx + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)dx \\
&\ge k^2\sigma^2 P(|X - \mu| \ge k\sigma)
\end{aligned}$$

$k^2\sigma^2$ 으로 양변을 나누면 $P(|X - \mu| \ge k\sigma) \le 1/k^2$. Q.E.D.

EXAMPLE 3-23 포아송 분포 이용

확률변수 X 는 평균 20, 표준편차 2를 따른다고 가정하자.

(1) 확률밀도함수를 모를 때 $P(16 < X < 24)$ 의 확률을 계산하시오.

(2) 확률밀도함수가 좌우 대칭일 경우 $P(16 < X < 24)$ 의 확률을 계산하시오.

0.75 / 0.95

Tchebysheff Inequality (체비세프 부등식)에 의해

$P(16 < X < 24) = P(-4 <| X - 20 |< 4) = P(-2*2 <| X - 20 |< 2*2) \ge 1 - \frac{1}{2^2} = 0.75$ 그러므로 0.75이다.

좌우 대칭이면 Empirical Rule에 의해 $P(-2*2 <| X - 20 |< 2*2)$, 즉 $\pm 2\sigma$ 에는 95%가 있으므로 0.95이다.