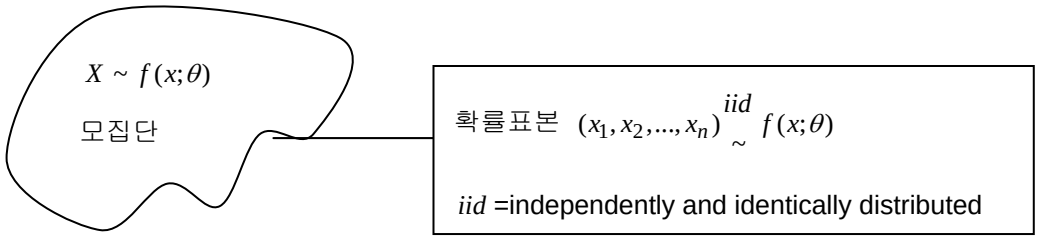
Chapter 3. 이산형 확률변수

3.1 정의(Definition)

확률변수 X 가 가질 수 있는(서로 다른) 값이 유한(finite)이거나 셀 수 있는(countable) 경우 이를 이산형(discrete)이라 한다. 서로 다른 값이 무한인 경우는 연속(continuous)이라 한다. 여기서는 이산형 확률변수에 대한 확률함수에 대해 논의할 예정이다. 표본 데이터로부터 얻은 통계량의 의 확률함수는 모집단 특성치, 모수(parameter)에 대한 추론에 사용된다. 이산형 확률변수의 확률함수는 쉽게 계산할 수 있다. 그러나 연속형인 경우에는 히스토그램(histogram), 혹은 Polygon이 확률함수 예측에 사용된다.



용어 정의

○모집단(population) 관심의 대상이 되는 개체(people, organization, animals, plants or things)의 모임

○표본(sample) 모집단에 대한 정보를 얻기 위하여 모집단으로부터 추출한 일부 개체.

○모수(parameter) 모집단의 특성을 요약한 값, θ 으로 표현하며 모집단 평균(μ), 모집단 분산(σ^2)이 대표적인 모수이다.

○통계량(statistic) 표본 데이터로부터 계산된 값, 표본평균($\bar{x}$), 표본분산(s^2)이 대표적인 예이다.

○추정(estimate) 모수 값을 점(point)이나 구간(interval) 값으로 표현

$\hat{\theta}, (\hat{\theta}_L, \hat{\theta}_U)$. For example, $\hat{\theta} = \bar{x}$, $(\bar{x} - t(n-1;1-\alpha/2) * s/\sqrt{n}, \bar{x} + t(n-1;1-\alpha/2) * s/\sqrt{n})$

○가설검정(Hypothesis testing) 모수에 대한 통계적 가설(귀무가설, 대립가설)의 진위 여부를 통계량을 이용하여 판단하는 과정

3.2 확률함수(Probability Distribution)

$(X = x)$, 표본공간 S의 모든 원소(결과)에 실수 x을 대응시킨 규칙을 확률변수라 한다.

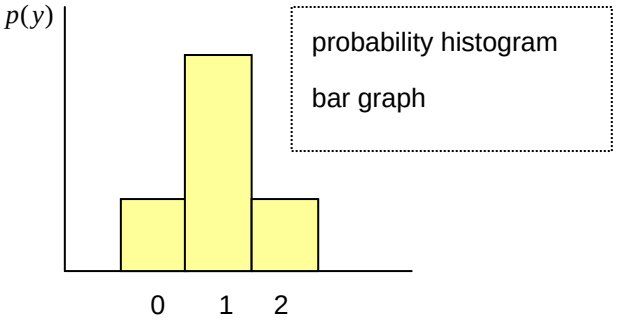
정의 확률변수 X가 임의의 값 x을 가질 확률을 $P(X = x)$ 혹은 $p(x)$으로 표현하고 표본공간 원소 중 x 값으로 대응된 원소의 확률의 합으로 정의한다.

정의 확률변수 X의 확률밀도함수(probability density function, pdf)는 확률변수 X가 가지는 값 x와 그에 대응하는 확률 $p(x)$을 그래프, 수식, 표 형태로 나타낸 것이다. x의 범위가 정의역(domain, X-축), 확률 $p(x)$가 치역(range, Y-축)인 함수이다. 이산형 확률변수의 확률밀도함수를 확률질량함수(probability mass function)이라 한다.

EXAMPLE 3-1 확률밀도함수

여자 3명, 남자 3명이 지원했다. 무작위로 2명을 선발할 때 선발될 남자의 수를 확률변수 X라 정의하자. 확률변수 X의 확률밀도함수를 구하시오.

y	$p(y)$
0	1/5
1	3/5
2	1/5



$$p(x) = \frac{\binom{3}{x}\binom{3}{2-x}}{\binom{6}{2}}, x = 0,1,2$$

EXAMPLE 3-2 확률밀도함수(2)

주머니에 구슬이 5개(1번-5번까지 숫자가 적혀 있음) 있다. 2개를 임의로 뽑았을 때

(1)확률변수 X을 **두 수 중 큰 수**라 정의할 때 확률밀도함수를 구하시오.

(2)확률변수 X을 **두 수의 합** 정의할 때 확률밀도함수를 구하시오.

정리(Theorem)

이산형 확률밀도함수는 다음 조건을 만족한다.

⑴ $0 \le p(x) \le 1,\ for\ all\ x$

⑵ $\sum_x p(x) = 1$

EXAMPLE 3-3 확률밀도함수 조건

아래 함수가 확률밀도함수가 되기 위한 상수 c 값을 구하시오.

⑴ $p(x) = cx,\ x = 1,2,...,10$

⑵ $p(x) = c(1/4)^x,\ x = 1,2,3,...$

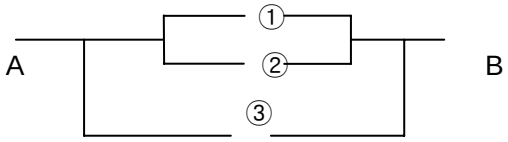
⑶ $p(x) = x/c,\ x = 1,2,3,...,n$

EXAMPLE 3-4 확률밀도함수 조건(2)

헌혈 자원자의 혈액을 조사하였더니 3명 중에 한 명은 O+, 15명 중에 1명은 O-이었다. 무작위로 3명을 선택하였을 때 확률변수 X 을 O+형을 가진 사람의 수, Y 을 O-형을 가진 사람의 수라 할 때 X, Y 의 확률밀도함수를 구하시오. 이를 결합(joint)확률밀도함수라 한다,

HOMEWORK #5-1

중계기가 작동할 확률은 0.9이고 서로 독립적으로 작동한다. 확률변수 X 을 작동하는 중계기 개수라 정의할 때 확률밀도함수를 구하시오.



HOMEWORK #5-2

주사위 X 는 숫자 0, 0, 0, 2, 2, 2, 주사위 Y 는 0, 1, 4, 5, 8, 9을 가지고 있다. 확률변수 W 을 $(X+Y)$ 라 정의할 때 확률변수 W 의 확률밀도함수를 구하시오.

3.3 기대값(The expected value)

확률밀도함수는 데이터의 분포(형태, 흩어진 정도)에 대한 정보는 주지만 요약 값(모수나 통계량)에 대한 정보를 제공하지는 않는다. 확률밀도함수의 숫자적 요약을 기술 요약값(descriptive values)이라 하는데 평균(mean), 중앙값(median), 분산(variance), 범위(range) 등이 그 예이다. 모집단의 기술 요약값을 모수, 표본인 경우에는 통계량이라 한다.

정의 확률변수 X 가 확률밀도함수 $p(x)$ 을 갖는다고 가정하자. 확률변수 X 의 기대값은 다음과 같이 정의한다. 이를 평균이라 한다.

$$E(X)=\sum_x xp(x)$$

$(X-E(X))^2$ 의 기대값을 확률변수 X 의 분산(variance)이라 하고 $V(X)=E[(X-E(X))^2]$ 표현한다. $V(X)$ 의 양의 제곱근을 확률변수 X 의 표준편차(standard deviation) 라 한다. 모집단의 경우 평균은 $\mu=E(X)$ 분산은 $\sigma^2=E(X-\mu)^2$ 으로 표현하고 표본 데이터의 경우 평균은 $\bar{X}=\sum x_i/n$, 분산은 $s^2=\sum(X-\bar{x})^2/n$ 으로 나타낸다.

확률변수 X 의 함수 $g(x)$ 의 기대값은 $E(g(X))=\sum_x g(x)p(x)$ 으로 정의한다.

정리(THEOREM)

⑴상수 에 대해 $E(c)=c$ 이 성립한다.

⑵ 상수 에 대해 $E[cg(X)]=cE[g(X)], E[cX]=cE[X]$ 이 성립한다.

⑶ $E[\sum_i^n a_iX_i]=\sum_{i=1}^n a_iE[X_i], E[\sum_i^n a_ig(X_i)]=\sum_{i=1}^n a_iE[g(X_i)]$, a_i for $i=1,2,...,n$ 는 상수.

⑷ $V(X)=E[(X-\mu)^2]=E(X^2)-\mu^2$

①위의 정리를 증명해 보자.

EXAMPLE 3-4 기대값

EXAMPLE 3-1에서 $E(X), E(1/X), E(X^2-1), V(X)$ 을 구해보자.

EXAMPLE 3-5 기대값 정리

새로운 기계 A 혹은 B를 사려고 한다. t 을 기계 작동 시간이라 하자. 기계 A의 일일 수리 회수 X_1의 평균, 분산 각각 $0.1t$ 이고 기계 B의 일일 수리 회수 X_2 는 평균, 분산 각각 $0.12t$ 을 갖는다. 기계 A의 일일 수리비용 함수는 $C_A = 10t + 30X_1^2$, 기계 B의 수리비용 함수는 $C_B = 8t + 30X_2^2$ 이다. 기계를 10시간 사용할 때, 20시간 사용할 때 비용을 최소화 하는 기계는?

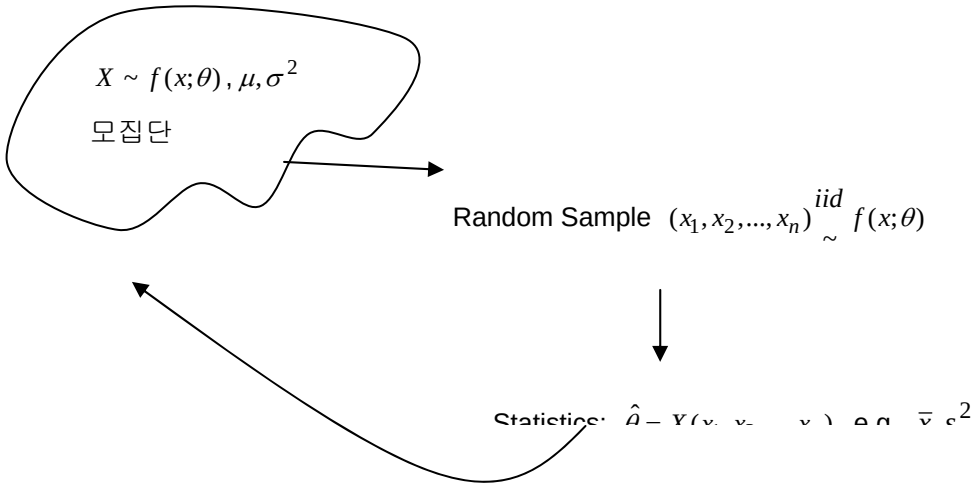
HOMEWORK #5-3

① a, b 는 상수, X 는 이산형 확률변수일 때 $E(aX+b) = aE(X)+b$ 이고 $V(aX+b) = a^2V(X)$ 임을 보이시오.

②확률변수 X 가 평균 μ, 분산 σ^2 을 가질 때 확률변수 $Z = \dfrac{X-\mu}{\sigma}$ (표준화. Standardized)의 평균과 분산을 구하시오.

HOMEWORK #5-4

세일즈맨이 하루에 고객 한 명을 만날 확률은 1/3, 두 명 만날 확률은 2/3이다. 만났을 때 판매에 성공할 확률은 0.1, 실패할 확률은 0.9이다. 판매에 성공하면 500,000원을 번다. 확률변수 X 을 일일 수입이라 정의할 때 $p(x), E(X)$ and $V(X)$ 을 구하시오.



기대치(Expected value) $E(X) = \sum_x xp(x)$ (평균), $V(X) = E[(X - E(X))^2]$ (분산)

•평균(mean): 관측치 크기의 중심

•분산(variance): 관측치들이 중심으로부터 얼마나 흩어져 있는지에 대한 측정

3.4 이항분포 (Binomial)

베르누이 시행(Bernoulli experiment)

•확률실험의 결과가 두 가지이다. (성공 success=1, 실패 fail=0)

•각 실험은 서로 독립(independent)이다.

•각 실험의 성공 확률은 p 으로 동일하다.

확률변수 X 을 베르누이 실행의 결과라 정의하자. 즉 성공이면 $X=1$, 실패면 $X=0$ 이다. 이 경우 확률변수 X 의 확률밀도함수(probability density function)은 다음과 같다. 이를 베르누이 분포(Bernoulli distribution)라 한다.

$$f(x) = p^x (1-p)^{1-x}, x = 0,1$$

Notation $X \sim Bernoulli(p)$

베르누이 분포에서 p 을 모수(parameter)라 한다. 여기서 모수의 의미는 그 값을 알면 확률밀도함수를 그릴 수 있다는 것이다. 즉 확률을 계산할 수 있다는 의미이다.

EXAMPLE 3-6 베르누이 분포 평균과 분산

확률변수 $X \sim Bernoulli(p)$ 의 평균과 분산이 p, pq 임을 보이시오.

$$E(X) = \sum xp(x) = \sum_{i=0}^{1} xp^x q^{1-x} = p\,, \quad E(X^2) = \sum x^2 p(x) = \sum_{i=0}^{1} x^2 p^x q^{1-x} = p$$

$$V(X) = E(X) - E(X^2) = pq$$

EXAMPLE 3-7 베르누이 분포와 이항 분포

학생들 중 감기 환자가 10%이다. 학생 3명을 임의로 선택하였을 때 확률변수 X 을 감기 걸린 학생 수라 정의할 때 X 의 확률밀도함수를 구하시오.

X	$P(X = x)$
0	
1	
2	
3	

정의(DEFINITION)

확률변수 X 을 n번의 베르누이 시행에서 성공 회수 정의하자. 확률변수 X 의 확률밀도함수는 다음과 같고 이를 이항분포(binomial distribution)라 한다.

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0,1,2,...,n \quad \text{--- ②}$$

Notation $X \sim Binomial(n, p)$

EXAMPLE 3-8 이항분포 PDF?

식 ②의 이항분포 PDF가 확률밀도함수임을 보이시오. 그리고 확률밀도함수의 의미는?

EXAMPLE 3-9 이항분포 이용

EXAMPLE 3-7에서 학생을 15명 임의 선택하였을 때 그 중 적어도 한 명이 환자일 확률을 구하시오.

0.5409

IN SAS

```
%macro loop1(n);
data one;
    do x=0 to &n;
    %do p=1 %to 9 %by 1;
        pr=&p/10;
        p&p=pdf('binomial',x,pr,&n);
    %end;
    output;
    end;
run;
proc print data=one;
    var x p1-p9;
run;
%mend loop1;
%loop1(15);
quit;
```

함수 PDF대신 CDF 사용하면 누적확률밀도함수를 얻는다.

$$F(x) = P(X \le x) = \sum_{t=0}^{x} \binom{n}{t} p^t (1-p)^{n-t}$$

x	p1	p2	p3	p4	p5	p6	p7	p8	p9
0	0.20589	0.03518	0.00475	0.00047	0.00003	0.00000	0.00000	0.00000	0.00000
1	0.34315	0.13194	0.03052	0.00470	0.00046	0.00002	0.00000	0.00000	0.00000
2	0.26690	0.23090	0.09156	0.02194	0.00320	0.00025	0.00001	0.00000	0.00000
3	0.12851	0.25014	0.17004	0.06339	0.01389	0.00165	0.00008	0.00000	0.00000
4	0.04284	0.18760	0.21862	0.12678	0.04166	0.00742	0.00058	0.00001	0.00000
5	0.01047	0.10318	0.20613	0.18594	0.09164	0.02449	0.00298	0.00010	0.00000
6	0.00194	0.04299	0.14724	0.20660	0.15274	0.06121	0.01159	0.00067	0.00000
7	0.00028	0.01382	0.08113	0.17708	0.19638	0.11806	0.03477	0.00345	0.00003
8	0.00003	0.00345	0.03477	0.11806	0.19638	0.17708	0.08113	0.01382	0.00028
9	0.00000	0.00067	0.01159	0.06121	0.15274	0.20660	0.14724	0.04299	0.00194
10	0.00000	0.00010	0.00298	0.02449	0.09164	0.18594	0.20613	0.10318	0.01047
11	0.00000	0.00001	0.00058	0.00742	0.04166	0.12678	0.21862	0.18760	0.04284
12	0.00000	0.00000	0.00008	0.00165	0.01389	0.06339	0.17004	0.25014	0.12851
13	0.00000	0.00000	0.00001	0.00025	0.00320	0.02194	0.09156	0.23090	0.26690
14	0.00000	0.00000	0.00000	0.00002	0.00046	0.00470	0.03052	0.13194	0.34315
15	0.00000	0.00000	0.00000	0.00000	0.00003	0.00047	0.00475	0.03518	0.20589

정리(THEOREM)

확률변수 $X \sim Binomial(n, p)$ 일 때 $\mu = E(X) = np,\ \sigma^2 = V(X) = npq$ 을 밝히시오.

PROOF

$$E(X) = \sum_x xp(x) = \sum_{x=0}^{n} x\binom{n}{x} p^x q^{n-x}$$

$$E(X) = \sum_{x=1}^{n} \frac{n!}{(n-x)!(x-1)!} p^x q^{n-x} (\because x = 0 \to xp(x) = 0)$$
$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(n-x)!(x-1)!} p^{x-1} q^{n-x} (z = x-1)$$
$$= np$$

$E(X(X-1)) = E(X^2) - E(X)$ 을 이용하여 $V(X)$ 을 계산하자. ($\because$ 이항분포의 경우 $E(X^2)$ 을 계산하는 것은 쉽지 않다)

$$E(X(X-1)) = n(n-1)p^2$$

그러므로 $E(X^2) = n(n-1)p^2 + np \Rightarrow V(X) = E(X^2) - E(X)^2 = npq$. Q.E.D.

HOMEWORK #6-1

확률변수 $X \sim Binomial(n, p)$ 이다. 확률변수 $(n - X)$ ("실패 개수")의 PDF를 구하시오. 그리고 어떤 분포를 따르는지 밝히시오.

HOMEWORK #6-2

헌혈 지원자의 80%는 헌혈 가능자라 하자.

⑴5명을 임의 추출했을 때 적어도 한 명이 헌혈 가능한 사람일 확률을 계산하시오.

⑵많아야 4명이 한혈 가능한 사람일 확률을 구하시오.

```
data one;
    do n=5 to 20;
    p=1-cdf('binomial',4,0.8,n);
    output;
    end;
run;
```

3.5 기하분포 (Geometric)

정의(DEFINITION)

베르누이 시행에서 첫 성공이 있기까지 실험(시행) 회수를 확률변수 X 라 정의하면 확률밀도함수는 다음과 같고 이를 기하분포라 한다.

$f(x) = p(1-p)^{x-1}, x = 1, 2, ...$

Notation $X \sim Geometric(p)$

IN SAS, PDF('GEOMETRIC',m,p); m=실패 회수, $m+1 = x$

EXAMPLE 3-10 기하분포 이용

기계 A가 1시간 동안 오작동 할 확률은 0.02이다. 이 기계가 두 시간 동안 오작동 없이 작동될 확률을 구하시오.

0.9604

정리(THEOREM)

$X \sim Geometric(p)$ 이면 $\mu = E(X) = 1/p,\ \sigma^2 = V(X) = (1-p)/p^2 = q/p^2$ where $q = (1-p)$ 이다.

PROOF

$$E(X) = \sum_x xp(x) = \sum_{x=1}^{\infty} xpq^{x-1} = p\sum_{x=1}^{\infty} xq^{x-1}$$

$\frac{d}{dq}(q^x) = xq^{x-1}$ 을 이용하면 $\frac{d}{dq}(\sum_{x=1}^{\infty} q^x) = \sum_{x=1}^{\infty} xq^{x-1}$ 임을 알 수 있다.

그러므로
$$E(X) = p\sum_{x=1}^{\infty} xq^{x-1} = p\frac{d}{dq}(\sum_{x=1}^{\infty} q^x)$$
$$= p\frac{d}{dq}(\frac{q}{1-q}) = \frac{1}{p}$$
. 분산은 숙제로 남겨둔다. Q.E.D.

EXAMPLE 3-11 기하분포 이용

은행에 고객이 임의의 1초에 창구를 찾을 확률은 0.1이고 고객이 창구를 찾는 사건은 서로 독립이라고 가정하자.

①3초에 첫 손님이 올 확률을 계산하시오.

②적어도 3초 안에는 첫 손님이 오지 않을 확률을 계산하시오.

③첫 손님이 오려면 몇 초나 기다려야 할까?

0.081 / 0.81 / 10

HOMEWORK #6-3

확률변수 $X \sim Geometric(p)$ 일 때 $P(X > a+b \mid X > a) = P(X > b)$ 임을 보이시오.

이를 기하분포의 **무기억성(memoryless property)**이라 한다.

HOMEWORK #6-4 (optional)

$X \sim Geometric(p)$ 일 때 $\sigma^2 = V(X) = (1-p)/p^2 = q/p^2$ where $q = (1-p)$ 임을 보이시오.

TIP $\dfrac{d^2}{dq^2}(\sum_{x=2}^{\infty} q^x)$ 을 이용하여 $E(X(X-1))$ 을 구한 후 $E(X^2)$ 을 얻으시오.

3.6 음이항 분포(Negative Binomial)

정의(DEFINITION)

베르누이 시행에서 r 번 성공할 때까지 시행하는 실험 회수를 확률변수 X 라 하면 확률밀도함수는 다음과 같고 이를 음이항분포라 한다.

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \ldots$$

Notation $X \sim NB(r, p)$

IN SAS, PDF('NEGBINOMIAL',x-r,p,r);

"Negative" Binomial 은 의미는 "이항분포의 역"을 의미한다. 이항분포는 n번 베르누이 시행에서 성공의 회수라면 음이항분포는 r번 성공할 때까지 시행의 회수에 대한 분포이다.

정리(THEOREM)

확률변수 $X \sim NB(r, p)$ 이면 평균과 분산은 다음과 같다.

$\mu = E(X) = r/p,\ \sigma^2 = V(X) = r(1-p)/p^2$

PROOF

확률변수 $X_1, X_1, ..., X_r$ 을 기하분포 $Geomeric(p)$ 라 하자.($X_i \underset{iid}{\sim} Geo(p)$) $\sum_{i=1}^{r} X_i$ 는 $NB(r, p)$ 을 따른다(이 부분은 5장에서 다루기로 한다).

평균 $E(Y) = E(\sum X_i) = \sum E(X_i) = \frac{r}{p}$ 이고 분산은 $V(Y) = V(\sum X_i) = \sum V(X_i) = r\frac{q}{p^2}$ 이다.

EXAMPLE 3-12 음이항 분포 이용

지질학 연구 결과 석유 탐사 중 석유 발견 확률은 0.2이다.
①3번째 탐사에서 처음 석유가 발견될 확률을 계산하시오.
②7번째 탐사에서 3번째 석유가 발견될 확률을 계산하시오.
③석유가 3개 발견될 때까지 시추를 계속하기로 하였다. 몇 번 정도 시추해야 하나?

0.128 / 0.049 /15

HOMEWORK #7-1

제품의 10%는 불량이다. 제품을 하나씩 검사하는 과정에서
(1) 첫 불량품이 2번째 검사에서 발견될 확률을 계산하시오.
(2) 5번째 시행에서 세 번째 불량 제품이 발견될 확률을 계산하시오.
(3) 5번째 시행 혹은 그 전 시행에서 세 번째 불량 제품이 발견될 확률을 계산하시오.

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.

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

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

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

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.

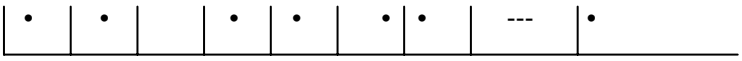
(1) $P(X=1)$ 일 확률을 구하시오.

(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.

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

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

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

위의 확률밀도함수를 포아송 분포라 한다.

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

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 $X_i \underset{\sim}{iid} Poisson(\lambda), i=1,2,...,n$ 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?

1.9×10^{-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

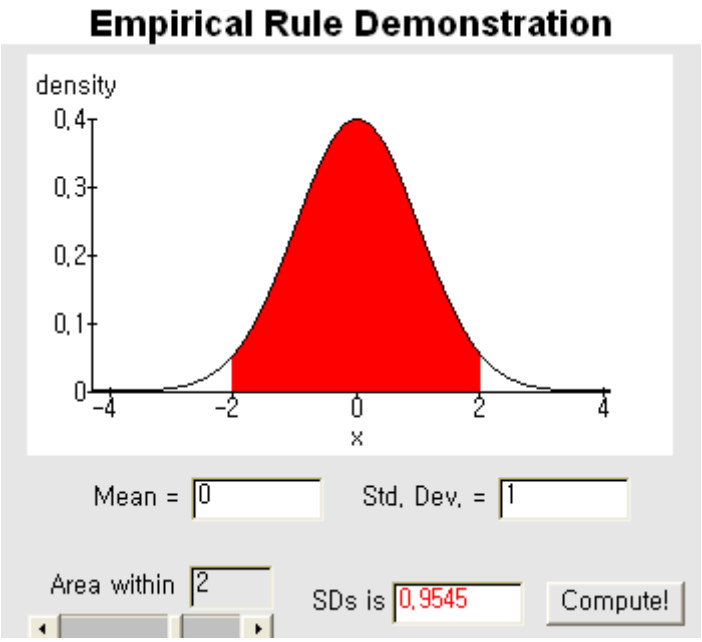
(1)이항분포를 이용하여 계산하시오 (2)포아송분포를 이용하여 계산하시오.

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)

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

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

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

HOMEWORK #7-5

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

HOMEWORK #7-6

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

⒜오전 9:00~9:30 사이에 정확하게 3명의 관람객이 박물관을 찾을 확률은?

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

⒞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

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

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

⒞ $Geometric(p)$ 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

아래 적률생성함수를 갖는 확률분포함수를 얻으시오. (확률분포함수 참고하시오)

⒜ $M(t) = [(1/3)e^t + (2/3)]^5$. ⒝ $M(t) = \frac{e^t}{2-e^t}$ ⒞ $M(t) = e^{2(e^t-1)}$

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.

$$P(|X-\mu| < k\sigma) \geq 1-\frac{1}{k^2} \text{ or } P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

PROOF

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

$k^2\sigma^2$ 으로 양변을 나누면 $P(|X-\mu| \geq k\sigma) \leq 1/k^2$. 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

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