2022학년도 2학기

경영 통계학

담당교수: 백수정

학습목표

1. 이산확률변수와 확률분포를 정의한다.
2. 기댓값과 분산 개념을 이용하여 문제를 풀 수 있다.
3. 이산균등분포를 정의하고 응용할 수 있다.
4. 엑셀, 관련 공식, 확률분포표를 이용하여 이항분포의 확률을 구할 수 있다.
5. 엑셀, 관련 공식, 확률분포표를 이용하여 포아송의 확률을 구할 수 있다
6. 엑셀을 이용하여 초기하분포의 확률을 구할 수 있다.

이산확률변수(discrete random variable)

- 확률변수(random variable)는 확률실험 표본공간(sample space)의 각 결과(outcome)에 숫자를 배정하는 규칙이나 함수를 의미
- 일반적으로 확률변수를 표기할 때 대문자 X를 사용하는 반면에, X의 구체적인 값에 대해서 소문자(예를 들어, x1)를 사용
- 이산확률변수(discrete random variable)는 셀 수 있는 값으로 구성되어 있음
- 어떤 확률변수는 분명하게 상한(40명이 한 반을 이루는 경우 결석한 학생의 수)을 가지고 있는 반면에 어떤 확률변수는 상한을 가지고 있지 않음(주어진 1시간 동안 당신이 받는 문자메시지의 수)

이산확률변수(discrete random variable)

의사결정 문제	이산확률변수
옥스나드 대학 MBA 프로그램의 신입생 정원은 65명이다. 작년에 합격한 학생 중 75%가 등록하였다. 80명의 학생을 합격시켰을 때 이 중 65명이 등록할 확률은?	X = 실제로 등록한 MBA 프로그램의 학생 수 (X = 0, 1, 2,⋯,80)
오전 9시부터 12시까지의 작업시간 동안 인터넷 쇼핑몰 주문센터 직원은 1분당 5건의 주문을 처리할 수 있다. 그리고 평균 주문량은 1분당 3.5건이다. 주어진 1분 동안 주문량이 5건 이상 될 확률은?	X = 주어진 1분 동안 인터넷 쇼핑몰에 걸려 오는 주문전화의 수 (X = 0, 1, 2,⋯)
어떤 강판 제조업자로부터 제공되는 강판은 길이 1미터 당 평균적으로 0.01개의 불량품이 생길 수 있다. 토요타 자동차는 500미터의 강판에서 10개 이상이 불량이면 그 주문을 거절한다. 임의의 주문이 거절될 확률은 얼마인가?	X = 강판 500미터에서 불량품의 개수 (X = 0, 1, 2,⋯)

확률분포(discrete probability distribution)

- 이산확률분포(discrete probability distribution)는 이산확률변수 X의 각 값에 확률을 대응시킴. 따라서 확률분포는 확률법칙을 반드시 따라야 함

- X가 n개의 값 $x_1, x_2, \ldots, x_n,$ 을 가지면,

$0 \le P(x_i) \le 1$ (확률변수 X의 각 값에 대한 확률)

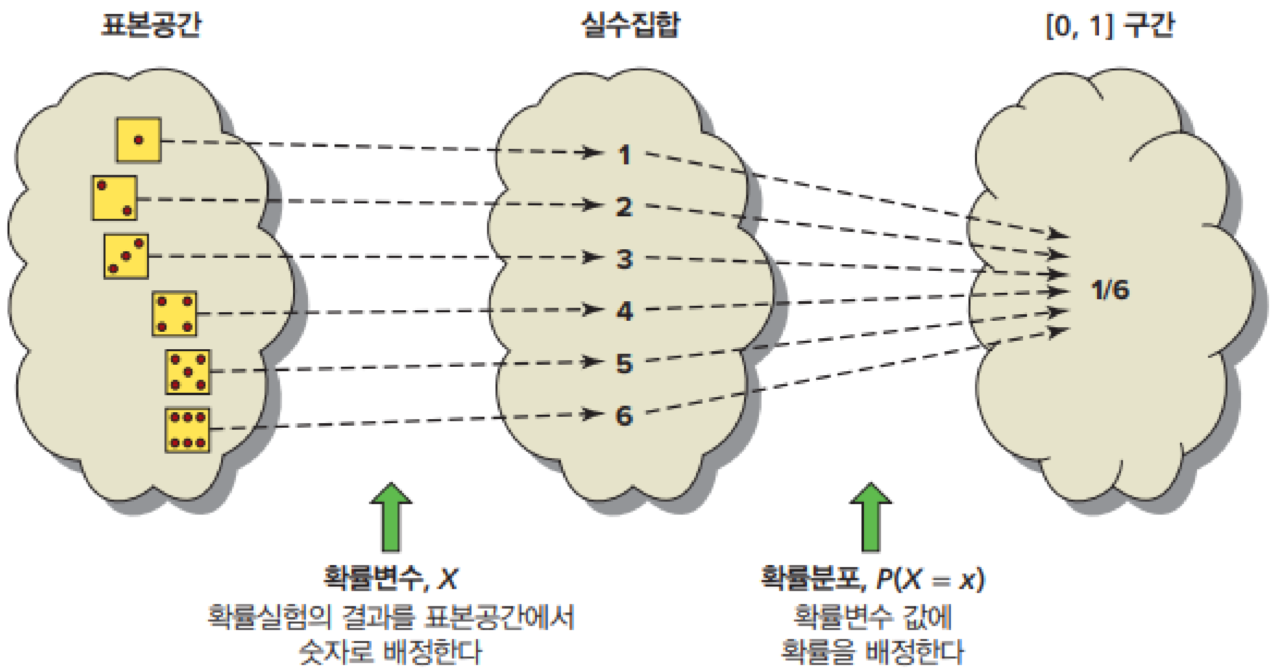
$\sum_{i=1}^{n} P(x_i) = 1$ (X의 모든 값에 대한 확률 합)

확률분포(discrete probability distribution)

- 이산확률분포는 수학에서 함수의 규칙을 따름
- 하나 이상의 결과가 있는 표본공간(sample space)은 같은 숫자로 대응될 수 있으나 하나의 사건 결과를 2개의 숫자로 대응할 수는 없음
- 유사하게 하나 이상의 확률변수 값은 같은 확률로 대응될 수 있지만, 하나의 확률변수 값은 2개 이상의 확률을 동시에 가질 수 없음
- 또한 확률의 합은 반드시 1이 되어야 함

확률분포(discrete probability distribution)

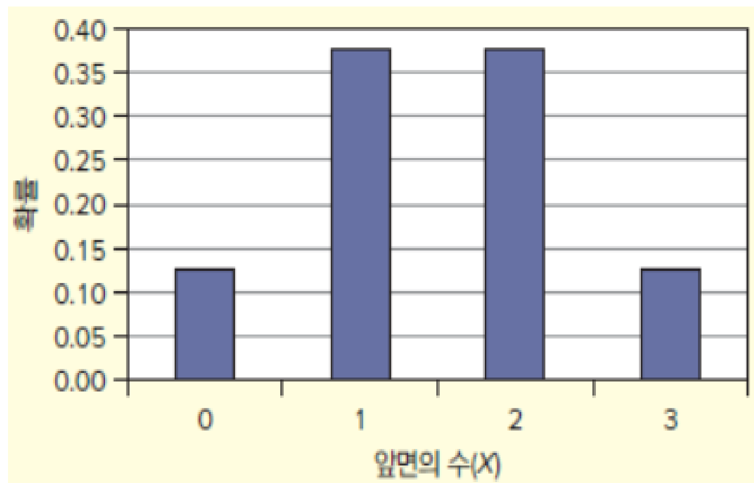
- 아래 그림은 주사위 1개를 던지는 경우에 표본공간, 확률변수 그리고 확률분포함수들의 관계를 보여줌



예제: 동전 던지기

- 동전 1개를 세 번 던질 때, 표본공간은 같은 확률을 가지는 8개 단순사건으로 구성됨(HHH,HHT, HTH, THH, HTT, THT, TTH, TTT). X를 앞면의 수로 정의하면 X는 아래 표와 그림에 제시된 확률분포를 가지는 확률변수

가능한 사건	x	$P(x)$
TTT	0	1/8
HTT, THT, TTH	1	3/8
HHT, HTH, THH	2	3/8
HHH	3	1/8
합계		1

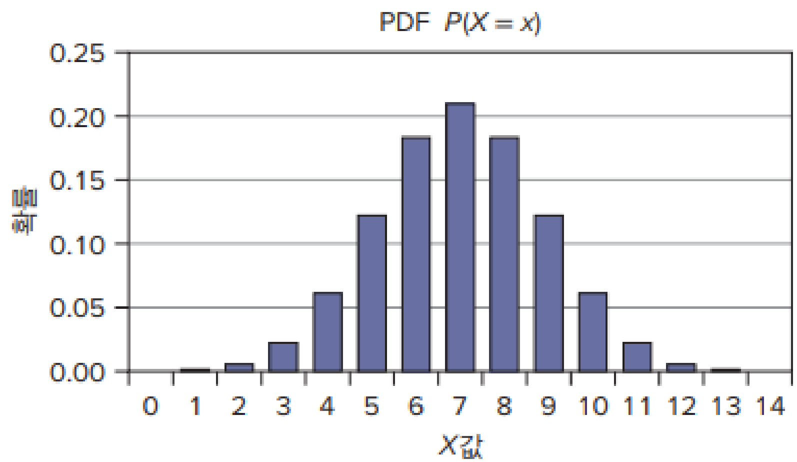


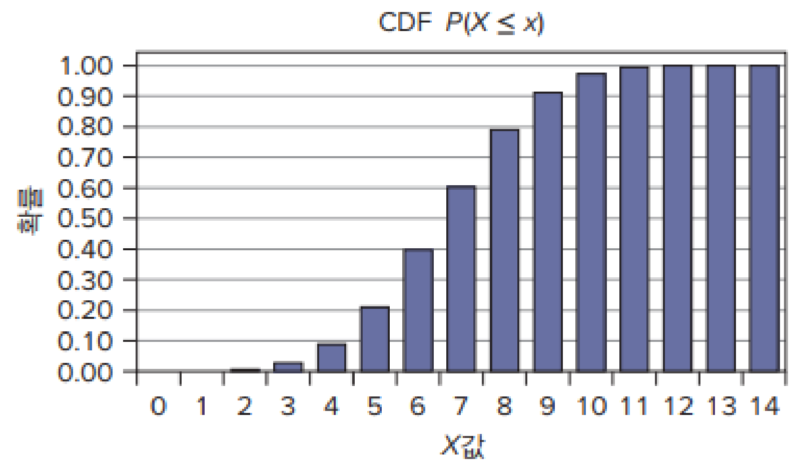
- X의 값이 모두 같은 확률을 가지는 것은 아님. 위 예제에서 X = 1과 X = 2는 X = 0과 X =3에 비해 그 가능성이 더 큼. 여느 확률분포와 마찬가지로 모든 확률을 더하면 1이 됨

PDF와 CDF는 무엇인가?

- 이산확률분포함수(probability distribution function, PDF): 각각의 X-값에 대해 확률을 나타냄
- 누적확률분포함수(cumulative distribution function, CDF): 확률의 누적 합계를 나타내며, X의 가장 작은 값에서 가장 큰 값까지 더하면 점점 1에 가까워 짐

PDF와 CDF 그래프





기댓값

- 이산확률 변수 기댓값 E(X)는 각 확률에 의해 가중평가된 X-값들의 합
- E(X)는 중심(center)을 측정하는 수단
- X가 n개의 값을 가지고 있다면, 기댓값은 다음과 같음

$$E(X) = \mu = \sum_{i=1}^{N} x_i P(x_i)$$

예제: 수리요청 전화

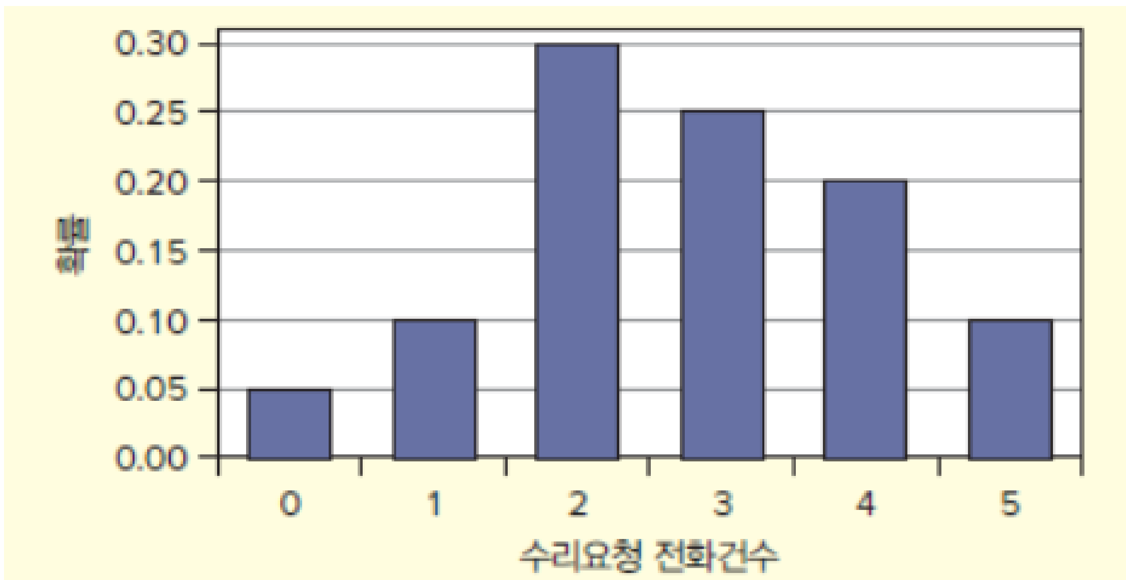
- 일요일에 에이스 가전제품 수리점에 걸려오는 수리요청 전화는 다음의 표와 같다. 다른 확률분포와 마찬가지로 확률의 합은 1이 된다.

X	$P(x)$	$xP(x)$
0	0.05	0.00
1	0.10	0.10
2	0.30	0.60
3	0.25	0.75
4	0.20	0.80
5	0.10	0.50
합계	1.00	2.75

$$E(X) = \mu = 0(0.05) + 1(0.10) + 2(0.30) + 3(0.25) + 4(0.20) + 5(0.10) = 2.75$$

예제: 수리요청 전화

- 아래 그림에 나타난 확률분포는 평균 μ = 2.75 주위에서 대칭이 아니라는 것을 알 수 있다. 그러나 평균 μ = 2.75는 여전히 균형점이다.



- E(X)는 평균이며, 반드시 관찰된 확률변수 값일 필요는 없다.

분산(variance)과 표준편차(standard deviation)

- 이산확률변수의 분산(variance) Var(X)는 각 X값과 기댓값 간의 편차를 제곱하여 확률 가 중치로 평균한 값. X가 n개의 값을 가지고 있다면 분산은 다음과 같음

$$\mathrm{Var}(X) = \sigma^2 = \sum_{i=1}^{N} [x_i - \mu]^2 P(x_i)$$

- 분산을 표기하기 위해 σ2과 Var(X)를 교환하여 사용
- 표준편차(standard deviation)는 분산의 제곱근 값이며 σ로 표기함

$$\sigma = \sqrt{\sigma^2} = \sqrt{\mathrm{var}(X)}$$

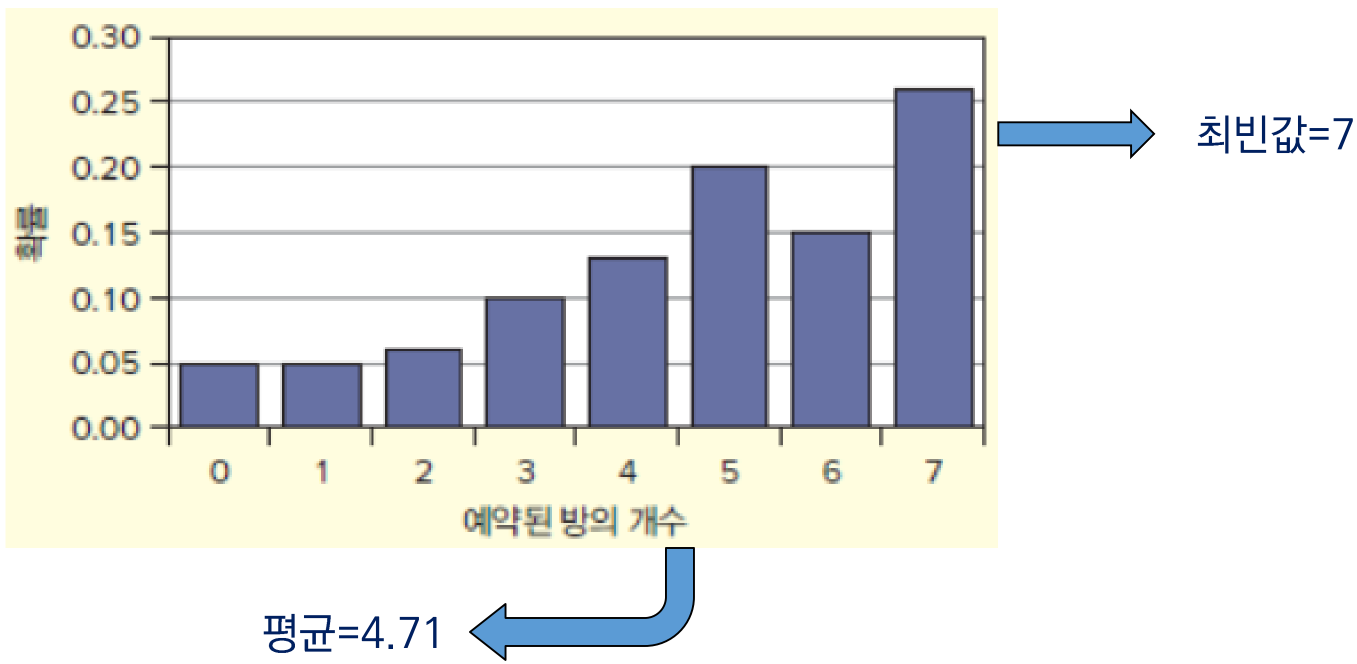
예제: 숙박 예약

- 베이 스트리트 모텔은 캘리포니아 산타테레사 지역에 있다. 객실에 대한 수요는 관광객이 많은 2월에 집중된다. 그러나 지금까지 경험으로 보아 객실 수요는 변동이 매우 심하다. 2월 중객실 수요는 아래 표와 같다.
  여기서 X = 예약된 객실의 수(X = 0, 1, 2, . . . , 7)이다. 아래 표는E(X)와 Var(X)를 계산하는 과정을 자세히 보여준다.

x	$P(x)$	$xP(x)$	$x-\mu$	$[x-\mu]^2$	$[x-\mu]^2 P(x)$
0	.05	0.00	– 4.71	22.1841	1.109205
1	.05	0.05	– 3.71	13.7641	0.688205
2	.06	0.12	– 2.71	7.3441	0.440646
3	.10	0.30	– 1.71	2.9241	0.292410
4	.13	0.52	– 0.71	0.5041	0.065533
5	.20	1.00	+ 0.29	0.0841	0.016820
6	.15	0.90	+ 1.29	1.6641	0.249615
7	.26	1.82	+ 2.29	5.2441	1.363466
합계	1.00	$\mu = 4.71$			$\sigma^2 = 4.225900$

예제: 숙박 예약

- 확률분포는 왼쪽으로 치우친 모양이며 양봉(bimodal) 모형이다. 최빈값(가장 많이 나타나는 값)은 7이지만, 2월에 예약되는 방의 평균 개수는 4.71개이다. 표준편차 2.06은 평균 주위에서 상당한 변동이 존재한다는 것을 보여준다. 아래 그림에서 확인할 수 있다.



이산균등분포(uniform distribution)

- 이산균등분포(uniform distribution)는 가장 단순한 이산확률분포의 하나이며, 이 분포의 확률변수는 a부터 b까지의 연속적 정수값을 가질 수 있음
- 분포의 전체적인 모양은 단지 2개의 모수 a와 b에 의존함
- 각 값이 나올 확률은 서로 같음

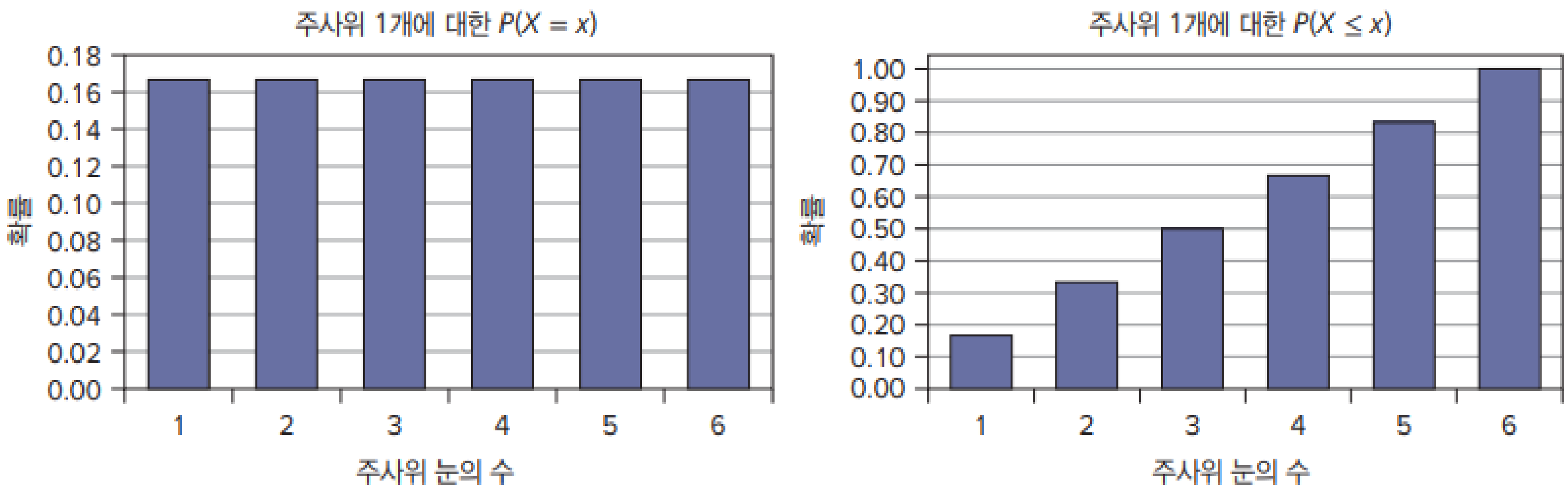
이산균등분포(uniform distribution)의 특징

모수	a = 하한 b = 상한
PDF	$P(X = x) = \dfrac{1}{b - a + 1}$
CDF	$P(X \le x) = \dfrac{x - a + 1}{b - a + 1}$
정의역	$x = a, a + 1, a + 2, \ldots, b$
평균	$\dfrac{a + b}{2}$
표준편차	$\sqrt{\dfrac{[(b - a) + 1]^2 - 1}{12}}$
엑셀에서 난수 생성	=RANDBETWEEN(a, b)
참고	무작위 정수나 다른 분포를 만들어 내는 데 사용된다.

예제: 주사위 1개를 던지는 경우

- 주사위 1개를 던지는 경우, 즉 6개의 정수값(1, 2, 3, 4, 5, 6)이 나올 확률이 동일한 이산균등분포를 형성함

주사위 1개를 던지는 경우 PDF와 CDF



예제: 주사위 1개를 던지는 경우

- 평균값 3.5는 반드시 1과 6 사이의 중간점이 되어야 한다는 것을 알 수 있음
  그러나, 표준편차의 경우에는 위의 공식을 이용하지 않고는 그 값을 예상하는 것이 쉽지 않음

평균: $\mu = \dfrac{a+b}{2} = \dfrac{1+6}{2} = 3.5$

표준편차: $\sigma = \sqrt{\dfrac{[(b-a)+1]^2 - 1}{12}} = \sqrt{\dfrac{[(6-1)+1]^2 - 1}{12}} = 1.708$

베르누이 실험(Bernoulli experiment)

- 두 가지 결과를 가진 확률실험
- 확률변수를 생성하기 위해서 우리는 임의로 하나의 결과를 성공(success: 이는 X = 1로 표기)으로 정의하고 다른 하나를 실패(failure: X = 0으로 표기)로 정의함
- 성공확률은 π로 표기하고, 실패확률은 1 − π 로 표기하며, 두 확률을 합하면 1이 됨

$$P(0)+P(1)=(1-\pi)+\pi=1.$$

- 성공확률 π 는 모든 시행에 있어 일정하게 유지됨

베르누이 실험(Bernoulli experiment)

- 아래 표에 있는 예는 성공(X = 1)이 꼭 바람직한 상황을 나타내는 것은 아님을 보여준다. 야금학자(metallurgist)는 금속의 피로점에 대한 징후를 찾으려 하고, 기업의 회계감사를 담당하는 사람은 지출에서의 문제점을 찾으려고 한다. 은행의 대출심사자는 부실채권 여부에 관심이 있다. '성공'은 단지 관심 있는 사건이라고 볼 수 있다.

베르누이 실험	가능한 결과	성공확률
동전 던지기	1 = 앞면 0 = 뒷면	$\pi = 0.50$
제트터빈 엔진 점검	1 = 문제점 발견 0 = 문제점 없음	$\pi = 0.001$
휘발유 주유	1 = 신용카드 사용 0 = 그 외 수단으로 지불	$\pi = 0.78$
엑스레이 검사	1 = 양성반응 0 = 음성반응	$\pi = 0.0004$

이항분포(binomial distribution)

- 베르누이 실험이 n번 반복될 때 정의
- 각 베르누이 실험은 서로 독립적이고 따라서 성공확률 π는 각 실험에서 일정하게 유지됨
- 이항실험에서는 X = n번 시도 시 성공 횟수를 참조하므로 이항확률변수 X는 n개의 서로 독립적인 베르누이 확률변수의 합으로 나타냄

$$X = X_1 + X_2 + \ldots + X_n$$

- 성공횟수에 대한 이항확률 P(X = x)는 두 모수인 n과 π에 의해 결정됨

이항분포(binomial distribution)의 특징

모수	n = 시행횟수 π = 성공확률
PDF	$P(X=x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$
엑셀*(PDF)	=BINOM.DIST(x, n, π, 0)
엑셀*(CDF)	=BINOM.DIST(x, n, π, 1)
정의역	$x = 0, 1, 2, \ldots, n$
평균	$n\pi$
표준편차	$\sqrt{n\pi(1-\pi)}$
엑셀에서 난수 생성	=BINOM.INV(n, π, RAND()) 엑셀에서 Data Analysis Tools
참고	$\pi < 0.50$이면 왼쪽 꼬리분포이고, $\pi > 0.50$이면 오른쪽 꼬리분포이다. $\pi = 0.50$이면 대칭적인 분포이다.

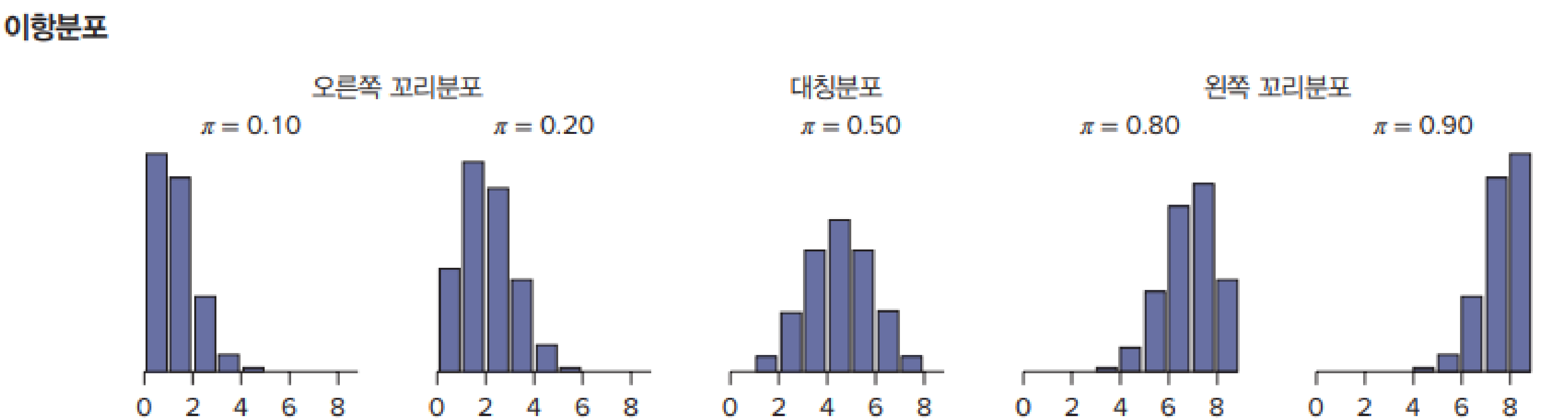
예제: 엔진오일 교환하는 자동차의 수

- 자동차 엔진오일 교환만 전문으로 하는 가게가 있다. 이런 유형의 사업에서는 엔진오일 교환시간이 오래 걸리지(late) 않는다고 고객이 인식하는 것이 중요하다. 그러므로 교환시간을 '오래 걸리는(late)'과 '오래 걸리지 않는(not late)' 의 두 가지 경우로 정의할 수 있고, 확률변수 X를 전체 서비스 차량 중에서 오래 걸리는 차량의 수로 정의한다. 그리고 각 차량의 교환시간이 오래 걸릴 확률은 모든 차에 대해서 같다고 가정한다. 또한 경험적으로 P(오래 걸리는 경우) = π = 0.10이라고 가정할 수 있다.
- 12대의 차량 중에서 정확히 2대가 오래 걸릴 확률을 알고자 하면 n = 12이고 우리는 P(X = 2)를 알고자 한다.

$$P(X=2)=\frac{12!}{2!(12-2)!}(.10)^{2}(1-10)^{12-2}=.2301$$

이항분포의 모양

- 이항분포는 π 〈 0.50이면 오른쪽 꼬리가 긴 분포이고 π 〉 0.50이면 왼쪽 꼬리분포가 되고, π = 0.50이면 정확히 대칭분포가 됨. 그러나 이러한 왜도(skewness)는 표본의 크기인 n이 커지면 그림 6.7에 나와 있듯이 π값과 상관없이 점점 줄어듬



이항분포 응용을 이해하기

- 이항분포는 다음과 같은 5가지 특징을 가지고 있음
  – 시행횟수(n)는 고정
  – 각 시행에서는 단지 두 가지 결과만 가능(성공과 실패)
  – 각 시행에서 성공확률(π)은 일정함
  – 각 시행은 서로 독립
  – 확률변수(X)는 n번 시행에서 성공횟수로 정의
- 성공확률 π를 알지 못하더라도 우리는 이항실험을 정의할 수 있음. 실제로 π값은 경험상 추정되는 값이지만 본 장에서는 주어진다고 가정함

이항확률 찾기

- 이항분포 공식을 이용
- 확률분포표 이용(부록A)
- 엑셀과 같은 통계패키지 이용
  (혹은 Mega Stat, Minitab이나 공학용 계산기 등 이용)

예제: 보험에 가입하지 않는 환자 수

- 평균적으로 볼 때 그린우드 종합병원 응급실 환자의 20%는 의료보험을 가지고 있지 않다. 4명의 환자를 무작위로 선택하였을 때 그 중 두 사람이 보험에 가입되어 있지 않을 확률은?
- X = 보험에 가입되지 않은 환자의 수 (“성공”)
- P(보험 미가입) = $\pi = .20$ (주어진 환자가 보험에 가입되지 않은 확률= 20%)
- P(보험 가입) = $1-\pi = .80$ (주어진 환자가 보험에 가입되어 있을 확률= 80%)
- N = 4명의 환자
- X의 범위는 0, 1, 2, 3, 4

$$\text{평균} = \mu = n\pi = (4)(0.20) = \text{환자 } 0.80\text{명}$$

$$\text{표준편차} = \sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(4)(0.20)(1-0.20)} = \text{환자 } 0.80\text{명}$$

예제: 보험에 가입하지 않는 환자 수

- 이항분포의 PDF는 표 6.8에 나와 있다. 우리는 엑셀의 이항공식(binomial formula)인 =BINOM.DIST(x, n, π, cumulative)를 이용하여 확률을 계산할 수 있다(여기서 PDF를 계산하고자 하면 cumulative는 0이고, CDF를 계산하고자 하면 cumulative 값은 1이 된다). 실제로 계산기를 이용하여 n = 4이고 π = 0.20인 이항확률 공식을 이용하여 계산할 수도 있다.

x	PDF $P(X = x)$	CDF $P(X \le x)$
0	0.4096	0.4096
1	0.4096	0.8192
2	0.1536	0.9728
3	0.0256	0.9984
4	0.0016	1.0000

- 두 환자가 보험에 가입되어 있지 않을 확률은 0.1536 혹은 15.36%

복합 사건(compound event)

- 부등호를 사용하는 사건
- 예: 4명의 환자 중 최대 2명의 환자가 보험 미가입자인 확률을 구해보자.
  이러한 확률은 P(X ≤ 2)라고 쓸 수 있다. PDF 값을 다음과 같이 더해서 구할 수 있다.

$$P(x \le 2) = P(X = 0) + P(X = 1) + P(X = 2) = .4096 + .4096 + .1536 = .9728$$

x	PDF $P(X = x)$	CDF $P(X \le x)$
0	0.4096	0.4096
1	0.4096	0.8192
2	0.1536	0.9728
3	0.0256	0.9984
4	0.0016	1.0000

복합 사건(compound event)

- 2명 미만의 환자가 보험 미가입자일 확률은?

$$P(X < 2) = P(0) + P(1) = .4096 + \ .4096 = .8192.$$

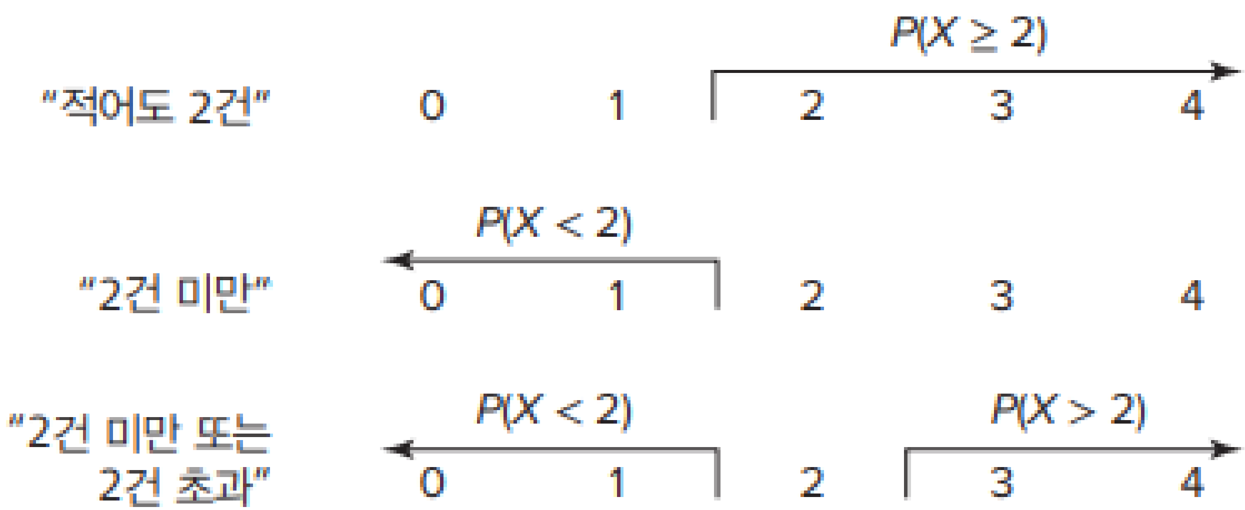
- 보험 가입자인 환자가 두 명을 초과하지 않을 확률은?

$$P(X \leq 2) = P(0) + P(1) + P(2) = .4096 + \ .4096 + .1536 = .9728$$

x	PDF $P(X = x)$	CDF $P(X \leq x)$
0	0.4096	0.4096
1	0.4096	0.8192
2	0.1536	0.9728
3	0.0256	0.9984
4	0.0016	1.0000

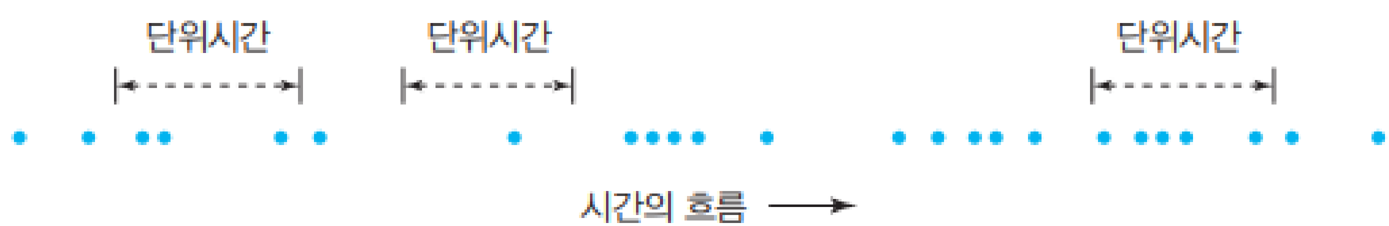
복합 사건(compound event)

- ‘초과(more than)’ 또는 ‘적어도(at least)’와 같은 문구를 해석하기 위해서는 아래와 같은 다이어그램을 그려보는 것이 도움이 됨



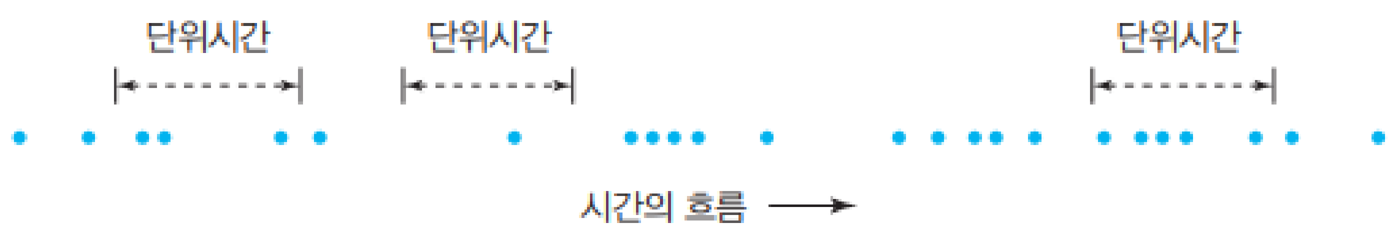
포아송 분포(Poisson distribution)

- 단위시간(예: 분, 시간) 또는 단위공간(예: 1제곱미터, 1마일)에서 무작위 발생횟수에 대한 확률분포
- 포아송분포가 적용되기 위해서는 아래 그림에서와 같이 연속된 시간이나 공간상에서 사건이 무작위적이고 독립적으로 발생하여야 함
- 대부분의 포아송분포 예는 단위시간 동안 도착횟수에 대한 모형에서 찾아볼 수 있으므로, 우리는 연속적인 상태를 ‘시간(time)’이라고 부르기로 함



포아송 분포(Poisson distribution)

- 각 점(·)은 관심있는 사건의 발생을 의미함
- X = 단위시간당 발생한 사건의 수로 정의
- X값은 단위시간이 어떻게 주어지느냐에 따라 변할 수 있는 확률변수임.
  아래 그림에서 사건이 발생하는 시간을 어디서 선택하느냐에 따라 X = 3 또는 X = 1 또는 X = 5의 값을 가짐



포아송 분포(Poisson distribution)

- 포아송 모형의 모든 특징은 평균인 λ(람다)에 의해서 결정됨
- 상수 e(자연로그의 밑수에 해당)는 대략 2.71828
- 포아송분포의 평균은 λ이고 표준편차는 평균에 대한 제곱근

- 포아송 공식의 장점: 단순성
- X는 상한이 없음. 즉 주어진 시간 동안 발생하는 사건의 횟수는 제한이 없음
- 그러나 포아송 확률은 X값이 증가함에 따라 확률이 0에 가까워지므로 실제 범위는 상당히 제한적임

포아송 분포(Poisson distribution)의 특징

모수	λ = 단위시간이나 공간에서 평균 도착횟수
PDF	$P(X=x)=\frac{\lambda^x e^{-\lambda}}{x!}$
엑셀* PDF	=POISSON.DIST(x, λ, 0)
엑셀* CDF	=POISSON.DIST(x, λ, 1)
정의역	$x = 0, 1, 2, \ldots$ (상한이 없다)
평균	λ
표준편차	$\sqrt{\lambda}$
참고	항상 오른쪽 꼬리분포, 그러나 λ가 커질수록 왜도가 작아진다.

예제: 은행 고객의 수

- 목요일 아침 9~10시에 옥스나드 대학 은행에 오는 고객은 평균적으로 1분에 1.7명이다. λ =1.7인 포아송 공식을 이용하면 PDF, 평균, 표준편차는 다음과 같이 도출한다.

$$\text{PDF: } P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(1.7)^x e^{-1.7}}{x!}$$

$$\text{평균}: \lambda = 1.7$$

$$\text{표준편차}: \sigma = \sqrt{\lambda} = \sqrt{1.7} = 1.304$$

예제: 은행 고객의 수

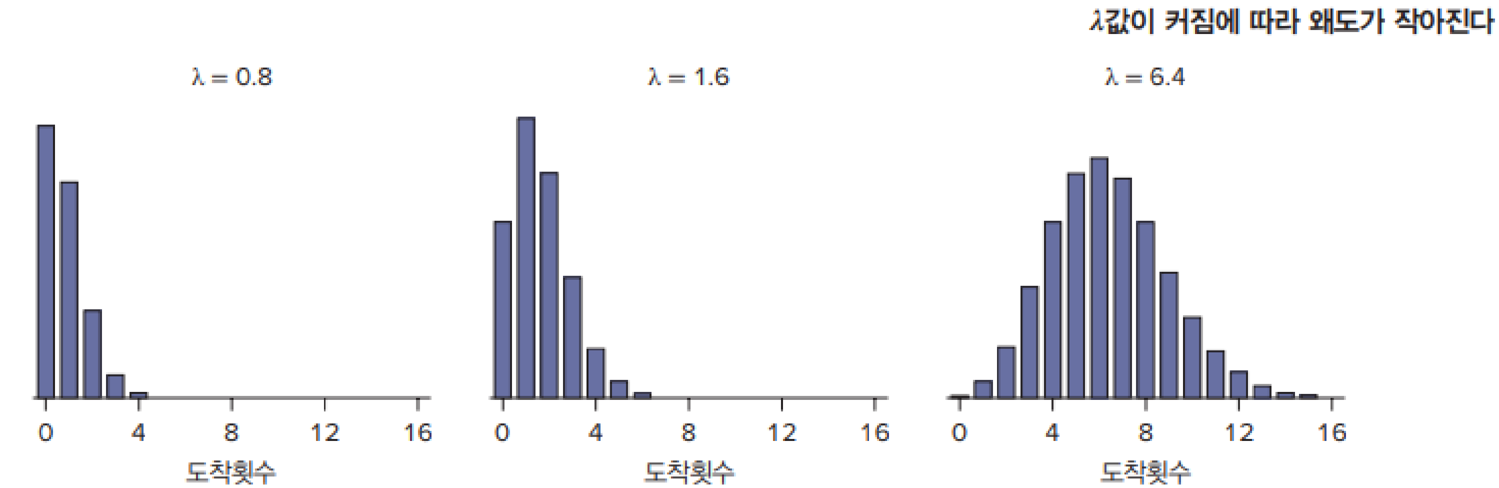
- 만약 1분에 은행 방문객이 정확히 3명이 되는 확률을 알고자 한다면 λ = 1.7, x = 3을 이용하여 계산한다.

$$P(X=3)=\frac{1.7^3 e^{-1.7}}{3!}=0.1496$$

- 공식을 이용하는 대신 엑셀 함수 =POISSON.DIST(3, 1.7,0)을 이용하여 같은 결과를 얻을 수 있다. 세 번째 모수 0은 P(X ≤ 3) 대신 (X = 3)을 계산하라는 의미이다.

포아송 분포의 응용

- 포아송분포의 네 가지 특징
  - 관심사건은 시간이나 공간상에서 무작위로 발생함
  - 평균 도착비율(λ)은 일정하게 유지됨
  - 사건발생(도착)은 서로 무관하게 이루어짐
  - 확률변수(X)는 일정한 시간 구간 내에서 발생한 사건 횟수



복합사건

- 개별 X의 확률이 더해짐으로써 누적 확률이 계산될 수 있음
- 예를 들어 주어진 1분 동안 2명 이하의 고객이 도착할 확률은 몇 개 사건의 합으로 계산됨

$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2) = .1827 + .3106 + .2640 = .7572$$

- 이 확률은 엑셀의 CDF 함수를 이용해서 계산 가능

$$P(X \le 2) = \text{POISSON.DIST}(2,\ 1.7,\ 1) = .7572$$

복합사건

- 적어도 3명이 도착할 사건(앞 슬라이드에서 나온 사건의 여사건)의 확률은 다음과 같이 계산

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - \text{POISSON.DIST}(2,\ 1.7,\ 1) = 1 - .7572 = .2428$$

- P(X ≥ 3)을 계산하고자 하는 경우 다음 페이지 그림에 노란색으로 표시된 것처럼 부록 B를 활용할 수 있음

복합사건

λ

X	1.6	1.7	1.8	1.9	2.0	2.1
0	.2019	.1827	.1653	.1496	.1353	.1225
1	.3230	.3106	.2975	.2842	.2707	.2572
2	.2584	.2640	.2678	.2700	.2707	.2700
3	.1378	.1496	.1607	.1710	.1804	.1890
4	.0551	.0636	.0723	.0812	.0902	.0992
5	.0176	*.0216*	.0260	.0309	.0361	.0417
6	.0047	.0061	.0078	.0098	.0120	.0146
7	.0011	.0015	.0020	.0027	.0034	.0044
8	.0002	.0003	.0005	.0006	.0009	.0011
9	–	.0001	.0001	.0001	.0002	.0003
10	–	–	–	–	–	.0001
11	–	–	–	–	–	–

포아송 분포 확률 찾기

- 포아송 공식을 이용
- 포아송 확률분포표(부록 B)활용
- 엑셀 이용하기
  (혹은 다른 통계 패키지인 Mega Stat, Minitab 등이나 공학용 계산기 이용)

초기하 분포(hypergeometric distribution)

- 모집단 N개로부터 비복원(without replacement) 추출을 한다는 것을 제외하고는 이항분포와 유사함
- 매 시행이 독립적이지 않고 성공확률이 시행할 때마다 일정하지 않음
- 초기하분포는 3개의 모수를 가지고 있음
  – N(모집단에 포함된 개체 수), n(표본에 포함된 개체 수), s(모집단에서 성공횟수)
- 초기하분포는 오른쪽 또는 왼쪽 꼬리 모양 분포이고 s/N = 0.50(즉 모집단에서 성공확률이 50%)이면 대칭분포가 됨

초기하 분포(hypergeometric distribution)의 특징

모수	N = 모집단에 속한 개체 수 n = 표본크기 s = 모집단에서 성공에 해당하는 개체 수
PDF	$P(X=x)=\dfrac{{}_sC_x\,{}_{N-s}C_{n-x}}{{}_NC_n}$
엑셀* PDF	=HYPGEOM.DIST(*x*, *n*, *s*, *N*, 0)
엑셀* CDF	=HYPGEOM.DIST(x, n, s, N, 1)
정의역	$\max(0, n-N+s) \le X \le \min(s, n)$
평균	$n\pi$, 여기서 $\pi = s/N$
표준편차	$\sqrt{n\pi(1-\pi)}\sqrt{\dfrac{N-n}{N-1}}$
참고	이항분포와 유사하지만 제한된 모집단에서 비복원 추출을 통해 표본을 얻는다. $n/N < 0.05$이면 $\pi = s/N$인 이항분포로 근사할 수 있다. $s/N = 0.50$이면 대칭분포가 된다.

예제: 손상된 아이패드

- 10개 아이패드 배송 중에 2개가 손상되었고 8개는 정상이다. 제품을 판매하는 마트에서는 3개 제품을 무작위로 뽑아서 불량인지 확인한다. 표본에서 불량인 아이패드 수는 확률변수 X이다. 다음과 같이 정리할 수 있다.
  - N = 10 (모집단에서 아이패드 숫자)
  - n = 3 (표본숫자)
  - s = 2 (배송 중 손상된 아이패드 숫자, 모집단에서 성공횟수)
  - N−s = 8 (모집단에서 정상 아이패드 숫자)
  - x = (표본에서 손상된 아이패드 숫자, 표본에서 성공횟수)
  - n−x = (표본에서 정상인 아이패드 수)

예제: 손상된 아이패드

- 위 문제를 이항분포로 이해할 수도 있다. 그러나 π가 일정하지 않다.
- 첫 번째 시도에서는 손상된 아이패드의 비율은 얼마인가? $\pi_1 = 2/10.$
- 두 번째 시도에서는 손상된 아이패드의 비율은 얼마인가?

$$\pi_2 = \frac{1}{9} \quad \text{(첫 번째 시행에서 손상된 아이패드가 뽑힌 경우)}$$

$$= \frac{2}{9} \quad \text{(첫 번째 시행에서 손상된 아이패드가 뽑히지 않은 경우)}$$

- 세 번째 시행 시, 두 번째 시행에 따라 다음과 같이 달라짐

$$\pi_3 = \frac{0}{8} \text{ or } = \frac{1}{8} \text{ or } = \frac{2}{8}.$$

예제: 손상된 아이패드

- 아이패드 예제에서 모집단에 손상된 아이패드 개수는 2이기 때문에 x의 가능한 값은 0, 1, 2이다. 확률은 다음과 같이 계산한다.

PDF 공식	엑셀 함수
$P(X=0)=\frac{{}_2C_0\,{}_8C_3}{{}_{10}C_3}=\frac{\left(\frac{2!}{0!2!}\right)\left(\frac{8!}{3!5!}\right)}{\left(\frac{10!}{3!7!}\right)}=\frac{56}{120}=\frac{7}{15}=0.4667$	=HYPGEOM.DIST(0, 3, 2, 10, 0)
$P(X=1)=\frac{{}_2C_1\,{}_8C_2}{{}_{10}C_3}=\frac{\left(\frac{2!}{1!1!}\right)\left(\frac{8!}{2!6!}\right)}{\left(\frac{10!}{3!7!}\right)}=\frac{56}{120}=\frac{7}{15}=0.4667$	=HYPGEOM.DIST(1, 3, 2, 10, 0)
$P(X=2)=\frac{{}_2C_2\,{}_8C_1}{{}_{10}C_3}=\frac{\left(\frac{2!}{2!0!}\right)\left(\frac{8!}{1!7!}\right)}{\left(\frac{10!}{3!7!}\right)}=\frac{8}{120}=\frac{1}{15}=0.0667$	=HYPGEOM.DIST(2, 3, 2, 10, 0)

초기하 분포의 응용

- 모집단의 숫자(N)가 제한적이고 성공횟수(s)가 알려져 있고 표본크기는 n이고 비복원 추출로 표본을 구함. 이러한 상황에서 성공확률은 매 시행마다 일정하지 않음

40대의 자동차가 캘리포니아 배기가스 규제에 적합한지 검사를 받는다. 32대는 적합판정을 받았고 8대는 부적합 판정을 받았다. 7대의 표본을 무작위로 뽑았다. 모든 자동차가 적합판정을 받을 확률은? 적어도 5대가 적합판정을 받을 확률은?
총기규제 당국은 총기 구매자에 대해 500가지 신원조회를 해야 한다. 50명의 신청자가중범죄로 기소된 적이 있다. 컴퓨터의 실수로 10명의 구매자가 신원조회 없이 승인되었다. 이들 중 중범죄자가 1명도 없을 확률은? 적어도 2명 이상 있을 확률은?
어느 의학실험실은 HIV 바이러스 보균 여부를 판단하기 위해 40명의 혈액검사를 실시 하였다. 8명이 실제로 HIV 보균자이다. 5개의 혈액샘플을 뽑았을 때, HIV 보균자가 없을 확률은?