제 4장 확률

우리는 일상생활에서 확률이라는 용어를 많이 접하게 된다. 확률(probability)는 한자어로 확실할 확(確), 비율 률(率)로 해석된다. 로또당첨확률, 야구 한국시리즈에서 특정 팀이 우승할 확률, 흡연자가 폐암에 걸릴 확률, 집값이 오를 확률 등 수없이 많은 확률들이 현대생활에서 사용되어지고 있다. 대부분의 일간신문에는 기상예보 면에 비나 눈이 올 확률을 제시하고 있다. '내일 비가 올 확률이 60%'라는 정보가 주어진다면 우리는 어떻게 이 정보를 유용하게 쓸 것인지, 60%의 의미는 정확히 무엇인지 아는 것은 현대생활을 영위하는 데 많은 도움을 줄 것이다. 확률은 불확실성을 길들이는 작업에서 중요한 개념이 된다. 몇 가지 예를 구체적으로 들어보자.

첫째 예: 다음 단락을 살펴보자.

감정 노동이란 원래 감정을 숨긴 채 배우가 연기를 하듯, 직업상 다른 얼굴 표정과 몸짓을 해야 하는 상황을 말합니다. 실제로 한 TV 프로그램에서 보듯이 서비스 업종에서 근무하는 직원들은 고객 만족 차원에서 무례하게 구는 고객에게도 웃는 얼굴로 대해야 합니다. 그러다 보니 서비스업 근로자들의 정신적 스트레스가 날로 심해져서 결국엔 우울증이나 대인 기피증으로 발전합니다. 그러나 감정 노동은 서비스 종사자에 국한되지 않습니다. 일반 직장인도 흔히 겪는 현상입니다. 감정 노동이 심해지게 되면, 구성원들은 심리적으로 탈진하게 되고, 조직성과도 하락하게 된다는데 문제의 심각성이 있습니다. 우선 구성원들은 업무 의욕이나 조직에 대한 충성심을 잃게 되기 쉽습니다. 특히 우수 인재들은 쓸데없는 감정에 체력을 소진하게 되고, 그 결과 자신의 역량을 제대로 발휘하지 못하고 조직을 떠나게 될 **확률**도 높아집니다.(송명석, 중앙일보 미디어마케팅연구소)

이 단락의 마지막 문장에서 구체적으로 몇 %라는 표현은 없어도 '확률'이라는 용어를 사용하였다.

둘째 예: 우리나라에서도 인기 만화를 원작으로 하는 드라마가 요즈음 인기가 있다. 다모, 궁, 어느 멋진 날, 다세포소녀 등. 외국순정만화 제목을 보면 재미있는 제목이 많은 데 '햇살가득 120%', '비올 확률 50%'같은 '확률'과 관련이 있는 만화제목이 눈에 띈다. 만화 '이키가미'(2006, 마제 모토로 저, 서현아 옮김, 학산출판사 간)는 1,000분의 1의 확률로 사람을 죽이는 바이러스를 사람들이 어릴 적에 투입해 성년이 되어 죽어가는 청년들의 죽음을 통하여 살아있는 사람들에게 경각심(앞으로 살 수 있는 시간은 24시간-당신을 무엇을 하고 싶으십니까?)을 준다는 내용으로 구성되어 있는 데 '1,000분의 1의 확률'이라는 표현이 나온다. '햇살가득 120%'이라는 표현은 문학적인 입장에서는 성립될 수 있으나 확률의 입장(확률은 0%보다 작을 수 없고 100보다 클 수 없다.)에서는 문제가 있다고 볼 수 있다. 이러한 또다른 예가 객석 2003년 8월호 편집장의 글-'어떤 추모'에서 보면 '120% 동의하면서'라는 표현이 나온다. 이러한 표현도 문학적인 입장에서는 성립될 수 있으나 확률의 입장에서는 문제가 있다고 볼 수 있다.

셋째 예: 왼손잡이에 대한 다음과 같은 단락을 보자.

중세에는 오른손잡이 전사가 칼싸움에서 상당히 유리했다. 이 들은 왼손으로 방패를 잡았기에 심장을 보호할 수 있었으므로 살아남아 후일 다시 싸울 수 있었다. 설사 칼에 찔리더라도 자손을 남길 수는 있었다. 19세기의 영국 작가 토마스 칼라일은 많은 사람들이 오른손잡이인 까닭은 바로 이 때문이라고 주장했다. 즉 왼손잡이들은 살아남아 자손을 번식할 기회가 없었다는 것이다. 그러나 이는 터무니없는 이론이다. 재치 있는 에세이 작가로 왼손잡이인 칼라일은 그가 못마땅해 하던 동시대의 인물 찰스 다윈을 비꼬기 위해 농조로 이런 이야기를 한 것 같다. 그러나 아직까지도 사람들이 오른손잡이 또는 왼손잡이가 되는 이유는 1871년 칼라일이 이 문제를 언급했을 당시와 마찬가지로 여전히 수수께끼로 남아 있다.(2000, 잉글리쉬 네티즌, 시사영어사)

우리나라에서 왼손잡이는 전 인구의 몇 %일까? 지난 25년간 미국대통령 6명 중 4명(제럴드 포드, 레널드 레이건, 조지 H. 부시(아버지 부시), 빌 클린턴)이 왼손잡이이다. 지미카터와 조지 W. 부시(아들 부시)만이 오른손잡이이다. 미국인 남성 중 15%, 여성 중 9%가 왼손잡이라고 한다. 왜 남성이 여성보다 왼손잡이가 많을까?

이러한 예들 외에 우리가 현대생활을 영위하는 데 확률이 쓰이는 실제 예들을 우리는 얼마든지 우리 주변에서 찾을 수 있다. 우리가 습득하는 지식들은 거의 대부분 불확실성을 내포하는 지식들이다. 이러한 불확실성을 계량화할 수 있다면 불확실성을 내포하는 지식들이 사용가능한 지식들이 된다(2003, Rao, 혼돈과 질서의 만남, 이재창 옮김, 나남출판사간). 앞에서 언급한 '내일 비가 올 확률이 60%'라는 일기예보를 예로 들어보자. 내일 우산을 가지고 외출해서 겪게 될 불편 때문에 원의 비용이 든다고 하고 내일 우산을 집에 두고 외출해서 비에 젖게 됨으로 발생되는 손실이 원이라 하자. 그러면 우리는 '내일 비가 올 확률이 60%'라는 정보를 앎으로 인하여 다음과 같은 표를 작성할 수 있게 된다.

의사결정	예상손실
우산을 가져간다 우산을 가져가지 않는다	

우리는 손실을 최소화해야하므로

이면 내일 우산을 가져간다
이면 내일 우산을 가져가지 않는다

라는 선택을 하게 된다. 이러한 선택이 가능한 이유는 '내일 비가 올 확률이 60%'라는 정보를 얻었기 때문이다. '내일 비가 올 확률'이 달라지면 우리의 선택도 달라지게 된다. 이렇듯이 불확실성을 계량화하는 수단으로서 확률이 유용하게 쓰이게 되는 것이다.

확률의 개념은 확률분포를 이해하는 데 가장 핵심적인 개념이 된다.

4.1 표본공간

통계적 시행(statistical experiment)이란 시행이 수행됐을 때 출현가능한 모든 결과들(outcomes) 중 오직 한 가지 나타나는 시행을 말하고 시행을 수행하기 전까지는 어떤 결과가 나올지 확실히 알 수 없는 시행을 가리킨다.

정의

표본공간(sample space) : 통계적 시행에서 출현가능한 모든 결과들의 집합

<예 4-1>

- 이산적 표본공간(discrete sample space): 표본 공간 상의 원소를 셀 수 있는 경우
  하나의 동전을 한 번 던 질 때:
  동전이 앞면이 나올 때까지 던질 때:

[참고] 동전에서 그림과 액면 문자(십원, 백원 등...)가 있는 쪽이 앞면이고 발행년도, 한국은행, 액면숫자(10, 100 등..)가 나온 면이 뒷면이다.

- 연속적 표본공간(continuous sample space): 표본 공간 상의 원소를 셀 수 없는 경우
  암환자 치료 후 생존시간: 』

정의

사건(event): 표본공간의 부분집합

<예 4-2>

- 하나의 주사위를 던졌을 때 눈의 수가 짝수가 나오는 사건:
- 52장의 트럼프카드(4개의 무늬 ♠, ♥, ♣, ◆가 있고 각 무늬 당 1(에이스)에서 10까지와 Jack, Queen, King 모두 13개의 카드가 있음.) 에서 랜덤하게 한 장을 꺼낼 때 에이스가 나오는 사건:
  ♠1, ♥1, ♣1, ◆1
- 주사위 두 개를 던졌을 때 눈의 수 합이 7이 나오는 사건:
  』

정의

단순사건(elementary event, simple event): 원소가 하나인 사건

<예 4-3>

- 하나의 동전을 던질 때 앞면이 나오는 사건: 』

정의

두 개의 사건 와 가 같은 원소가 없을 때, 즉 일 때 두 개의 사건 와 는 서로 배반인(mutually exclusive) 사건이라고 한다.

<예 4-4> 하나의 주사위를 던질 때 짝수의 눈이 나오는 사건과 홀수의 눈이 나오는 사건은 서로 배반인 사건이다. 』

4.2 확률

확률은 통계적 시행에서 어떤 특정 사건이 일어날 가능성을 측정하는 수치적인 측도이다. 이러한 확률을 정의하는 데 크게 3가지 접근 방법이 있다.

1. 확률의 고전적인 정의(classical definition, Laplace가 제창함.)

표본공간의 모든 원소가 일어날 가능성이 같다고 가정하면 사건 의 확률은 다음과 같이 정의된다.

P(A)=사건 A에 속하는 모든 원소의 개수/표본공간의 모든 원소의 개수

<예 4-4>

- 하나의 주사위를 던졌을 때 눈의 수가 짝수가 나오는 사건을 라 하자. 사건 의 확률은 P(A)=사건 A에 속하는 모든 원소의 개수/표본공간의 모든 원소의 개수=3/6=1/2이 된다.
- 로또 1등 당첨확률= 』

이러한 확률의 고전적인 정의 고등학교 수학과 교과과정에서는 수학적 확률(mathematical probability)이라고 칭하였다. 그런데 이러한 고전적인 정의의 맹점은 표본공간의 모든 원소가 일어날 가능성이 같다고 가정하는 데 있다. 다음과 같은 확률들을 알고 싶다고 하자.

- 자동차공장에서의 제품불량률
- 임의로 선택한 한 가족이 집을 소유하고 있을 확률
- 임의로 선택한 한 여성이 흡연자일 확률
- 찌그러진 동전을 던질 때 앞면이 나올 확률
- 임의로 선택한 한 사람이 SUV(Sport-Utility Vehicle)을 소유하고 있을 확률

이러한 확률은 확률의 고전적인 정의로는 계산할 수 없게 된다. 그래서 다음과 같은 상대도수(relative frequency)의 극한적인 개념으로 확률을 정의할 수 있다. 이러한 상대도수의 극

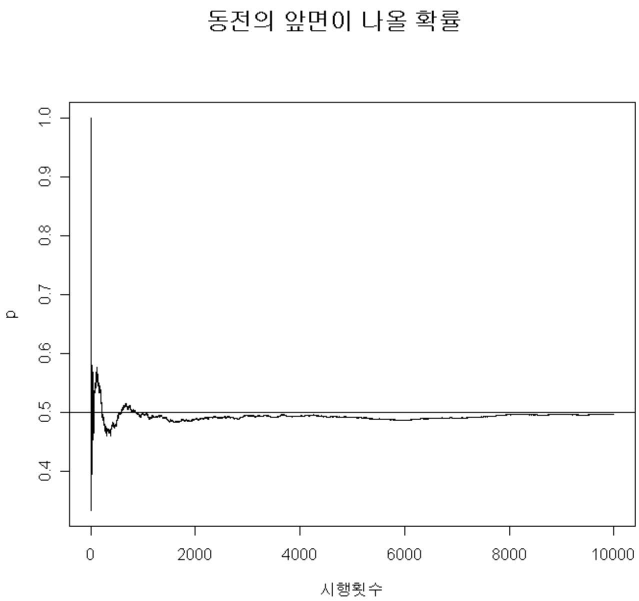
한적인 개념을 고등학교 수학과 교과과정에서는 통계적 확률(statistical probability)이라고 칭하였다.

2. 상대도수의 극한적인 개념으로서의 확률

여기서 은 시행횟수, 는 번의 시행 중 사건 가 일어나는 횟수를 나타낸다.
이러한 상대도수의 극한적인 개념으로 확률을 정의할 때 두 가지 사실을 주의하여야 한다.

(1) 상대도수에서 을 크게 하면 수렴값에 수렴하느냐(존재성, existence)
(2) 이 수렴값이 여러 개가 아니고 딱 하나이냐(일의성, uniqueness)

<예제 4-5> 다음 그림은 컴퓨터를 이용하여 동전을 10,000번 던져 동전의 앞면이 나올 확률이 0.5임을 보이는 그림이다. 시행회수 이 작을 때는 동전의 앞면이 나올 확률이 매우 불안정하게 흔들리나 이 커지면서 0.5에 수렴함을 알 수 있다.



』

확률의 고전적인 정의와 상대도수의 극한적인 개념으로서의 확률을 아우르는 정의가 다음과 같은 확률의 공리적인 정의이다.

3. 확률의 공리적인 정의(axiomatic definition, Kolmogorov가 제창함.)

표본공간 와 사건 에 대하여 다음 3 가지 공리를 만족할 때 P(A)를 사건 A의 확률이라 고 한다.

(공리 1)
(공리 2)
(공리 3) 서로 배반인 사건 에 대하여 다음 수식이 성립한다.

4.3 확률의 법칙

4.3.1 여사건의 확률

표본공간 에서 정의되는 사건 에 대하여 여사건(complementary event) 의 확률은 다음과 같다.

<예제 4-6> 1912년 4월 14일 밤 호화여객선 타이티닉호가 할리팩스 부근 북대서양에서 빙하와 부딪쳐 침몰하였다. 이 타이타닉호 참사는 워낙 유명한 사건이라 영화로도 여러 번 만들어졌다. 1997년 제임스 카메론 감독이 촬영한 영화가 최근판인 데 레오나르도 디카프리오가 하류계층(3등석) 남자주인공으로, 케이트 윈슬렛이 상류계층(1등석) 여자주인공으로 나와 이루어질 수 없는 사랑이야기가 주축이 되어 전개되었다. 이 타이타닉호 참사에 대한 다음과 같은 분할표(contingency table)를 이용하여

	계층				합계
	1등석	2등석	3등석	승무원	
생존	202	118	178	212	710
사망	123	167	528	673	1,491
합계	325	285	706	885	2,201

(a) 총 2,201명의 사람들 중 임의로 선택한 사람이 살아남을 확률을 구하라.
(b) 총 2,201명의 사람들 중 임의로 선택한 사람이 살아남을 확률을 구하라.

(풀이) (a) 총 2,201명의 사람들 중 임의로 선택한 사람이 살 사건을 라 하면
710/2,201=0.3226=32.26%
(b) 총 2,201명의 사람들 중 임의로 선택한 사람이 죽을 사건을 라 하면

1,491/2,201=0.6774=67.74%. 여사건의 확률 공식을 이용하여 구하면
=1-0.3226=0.6774=67.74%

4.3.2 확률의 덧셈정리

표본공간 에서 정의되는 두 개의 사건 와 에 대하여 합사건(union of events) 의 확률은 다음과 같다.

<예제 4-7>(<예제 4-6> 계속) 총 2,201명의 사람들 중 임의로 선택한 사람이 살아남거나 승무원일 확률을 구하라.

(풀이) 총 2,201명의 사람들 중 임의로 선택한 사람이 살 사건을 라 하고 승무원일 사건을 라 하면

』

4.4 조건부확률

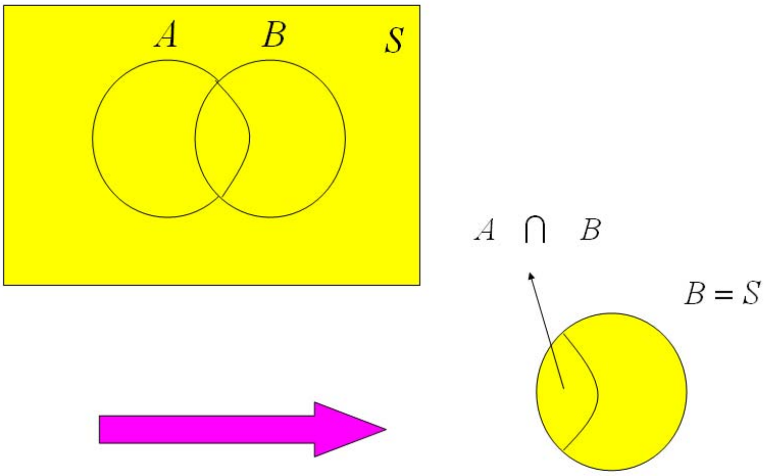
4.4.1 조건부확률

사건 가 주어진 조건 하에서 사건 가 일어날 조건부확률(conditional probability of event A given event B) 는 다음과 같이 정의된다.

정의

단, 이어야 한다.

위의 정의를 음미해 보기 위하여 벤다이어그램을 그려 보면



사건 가 주어진 조건 하에서 사건 가 일어나야 하므로 가 로 축소된다고 생각할 수 있으므로 사건 가 주어진 조건 하에서 사건 가 일어날 조건부확률은 중에서 가 차지하는 비율이라 볼 수 있으므로 로 정의하는 것이 타당하다 하겠다.

<예제 4-8>(<예제 4-6> 계속)

(a) 총 2,201명의 사람들 중 임의로 선택한 사람이 1등석 승객이면서 살아남을 확률을 구하라.

(b) 1등석 승객 중 살아남을 확률을 구하라.

(c) 살아남은 사람 중 1등석 승객일 확률을 구하라.

(d) 총 2,201명의 사람들 중 임의로 선택한 사람이 3등석 승객이면서 살아남을 확률을 구하라.

(e) 3등석 승객 중 살아남을 확률을 구하라.

(f) 살아남은 사람 중 3등석 승객일 확률을 구하라.

(g) (a)-(f)를 통하여 알 수 있는 사실은 무엇인가?

(풀이) (a) 총 2,201명의 사람들 중 임의로 선택한 사람이 1등석 승객인 사건을 이라하고 임의로 선택한 사람이 살 사건을 라 하면 202/2,201=0.0918=9.18%

(b) =202/325=0.6215=62.15%

(c) =202/710=0.2845=28.45%

(d) 총 2,201명의 사람들 중 임의로 선택한 사람이 3등석 승객인 사건을 이라하고 임의로 선택한 사람이 살 사건을 라 하면 178/2,201=0.0809=8.09%

(e) =178/706=0.2521=25.21%

(f) =178/710=0.2507=25.07%

(g) 우리가 제일 알고 싶은 정보는 ‘모든 계층이 공평하게 사망비율(또는 생존비율)이 같으냐’하는 것일 것이다. 즉, (b)와 (e)가 중요한 정보가 된다. 그런데 상류계층의 생존비율은 62.15%인 데 반하여 하류계층의 생존비율은 25.21%뿐 이어서 2.47배의 차이가 남을 알 수 있다. 공평하지 않았음을 알 수 있다.

4.4.2 확률의 곱셈정리

표본공간 에서 정의되는 두 개의 사건 와 에 대하여

가 성립한다. 단, 첫째 식에서는 이 성립되어야 하고 둘째 식에서는 이 성립되어야 한다.

이 곱셈정리를 일반화하면 다음과 같은 일반화된 곱셈정리가 성립한다.

표본공간 에서 정의되는 사건들 에 대하여

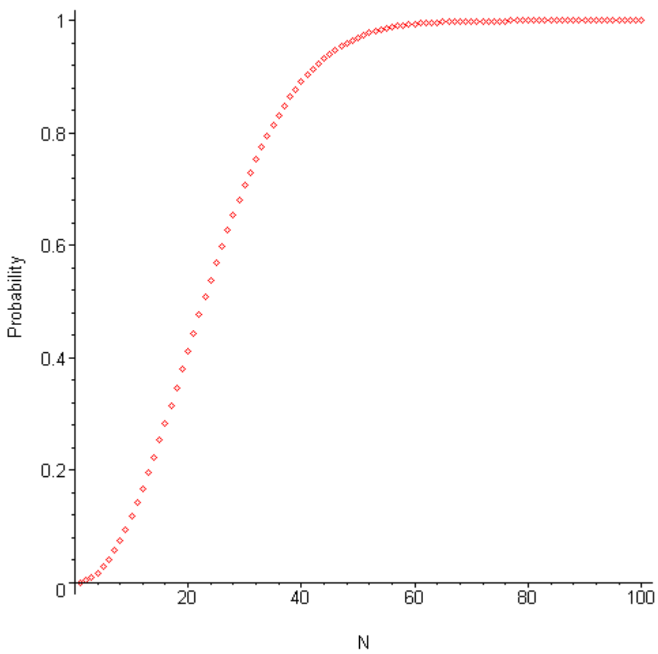
여기서 를 의미하고 을 의미한다.

<예제 4-9> 한 방에 모여 있는 사람의 수가 이라 하자. 한 방의 사람들 중 서로 같은 생일을 갖게 될 확률을 구하여라(1년은 365일로 가정함.).

(풀이) 우선 P[명이 같은 생일을 갖지 않는다]를 구하여보자. 첫 번째 사람이 365일 중 하루를 선택할 확률은 , 첫 번째 사람이 365일 중 하루를 선택하면 두 번째 사람은 나머지 364일 중 하나를 선택하여야 하므로 두 번째 사람이 하루를 선택할 확률은 , 이런 식으로 생각하면 일반화 곱셈정리를 이용하여 P[명이 같은 생일을 갖지 않는다]= 이 된다. 그러므로 우리가 구하여야 할 확률은 여사건의 확률공식을 이용하여 P[명 중 최소한 두 명이 같은 생일을 갖는다]=1-P[명이 같은 생일을 갖지 않는다]= 이 된다. 이 확률을 표로 작성하면 다음 표와 같다.

	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75
	0.027	0.117	0.253	0.411	0.569	0.706	0.814	0.891	0.941	0.970	0.986	0.994	0.998	0.999	0.9997

에 따른 확률 P[명 중 최소한 두 명이 같은 생일을 갖는다]의 변화를 그림으로 그리면 다음과 같다.



한 방에 23명이 모여 있으면 최소한 두 명이 같은 생일을 갖게 될 확률이 0.507로 0.5를 넘게 되고 한 방에 40명이 모여 있으면 최소한 두 명이 같은 생일을 갖게 될 확률이 0.891로 매우 큼을 알 수 있다. 』

4.5 독립사건

정의

표본공간 에서 정의되는 두 개의 사건 와 에 대하여

와 가 서로 독립(mutually independent)

이면 와 가 서로 종속(mutually dependent)이라고 한다. 위의

정의로부터 와 가 서로 독립이면 다음과 같은 성질이 성립한다.

단, 첫째 식에서는 이 성립되어야 하고 둘째 식에서는 이 성립되어야 한다.

[참고] ‘서로배반’과 ‘서로독립’은 다른 개념이다. 이고 일 때 ‘서로배반’이면 ‘서로종속’이 된다.

<예제 4-10>(<예제 4-6> 계속) 총 2,201명의 사람들 중 임의로 선택한 사람이 살 사건을 라 하고 승무원일 사건을 라 할 때 사건 와 사건 는 독립인가?

(풀이)

사건 와 사건 는 종속이다. 』

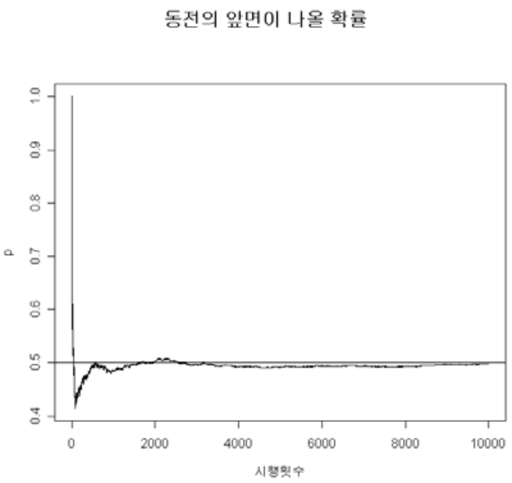
4.6 R 실습

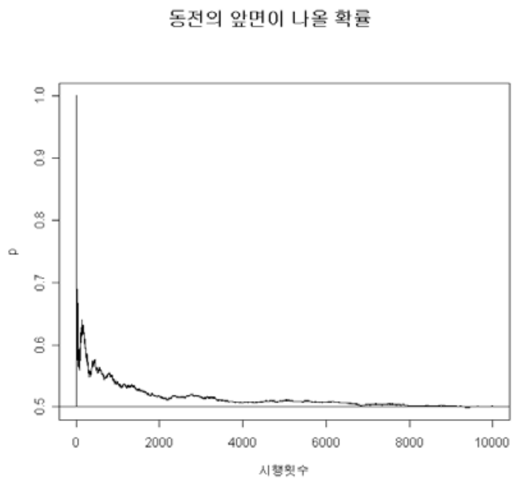
1. R에서의 난수를 이용하여 동전의 앞면이 나오는 확률을 구하여 보아라.

(풀이) 동전의 앞면이 나오는 확률을 구하기 위한 R 프로그램은 다음과 같다.(시행횟수를 10,000번으로 하였으나 조정이 가능함.)

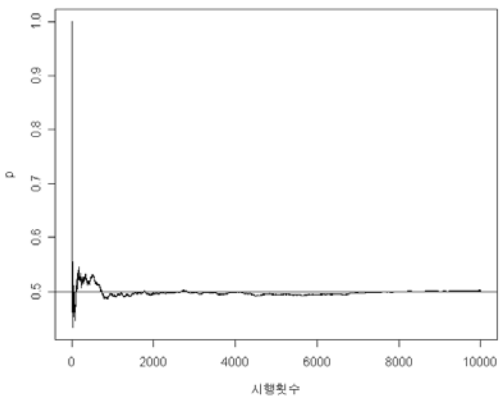
```
law.large.number=
function(n)
{
win.graph()
par(oma=c(0,0,5,0))
par(mfrow=c(1,1))
# 베르누이분포에서의 난수
x1=rbinom(n,1,0.5)
# 상대도수
xbar1=cumsum(x1)/1:n
plot(xbar1,ylab="p",xlab="시행횟수",type="l")
abline(0.5, 0)
mtext("동전의 앞면이 나올 확률",side=3,outer=T,cex=1.5)
par(mfrow=c(1,1))
}
law.large.number(10000)
```

동전을 10,000번 던져 동전의 앞면이 나오는 확률을 다섯 차례 구하여보니 다음과 같은 다섯 가지 그림이 나왔다. 첫 번째, 두 번째, 세 번째 그림에서는 0.5에 수렴하는 것이 보이나 네 번째나 다섯 번째 그림에서는 동전을 10,000번 던졌음에도 불구하고 0.5에 수렴하는지 알기가 쉽지 않다. 그만큼 동전의 앞면이 나오는 확률과 같은 간단한 확률도 컴퓨터로 구현하여 보는 것이 쉬울 때도 있으나 간단치 않을 때도 있다. 즉, 컴퓨터를 이용한 난수(random number)는 완전한 난수가 아니라 의사난수(pseudo-random number)이기 때문에 이러한 문제점이 발생하게 되는 것이다.

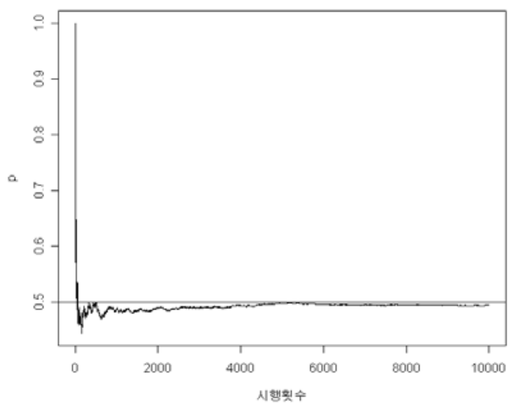




동전의 앞면이 나올 확률



동전의 앞면이 나올 확률



동전의 앞면이 나올 확률