확률(probability)이란 이러한 사건들이 일어날 수 있는 가능성(likelihood)을 0에서 1 사이의 숫자로 나타낸 척도다. 표본 공간 안에 있는 근원 사건들은 총합이 1인 확률을 일어날 수 있는 가능성에 비례하여 나눠 갖는다. 핵심!!! 기본!!! 확률 분포(distribution)에서 'distribute'은 총합이 1인 확률을 나눠준다는 의미다. 동전 던지기의 표본 공간에는 '앞면(H)' 혹은 '뒷면(T)'이라는 원소가 있으므로 1이라는 숫자를 두 원소에게 나눠준다. 공정한 동전(fair coin)이라면 앞면과 뒷면이 나올 가능성은 같으므로 각각 0.5라는 확률을 나눠 갖는다.

이렇게 확률을 사건들에게 나눠주는 방식에 따라서 불확실한 세상은 아주 다양한 모습을 보여준다. 게임도 다르지 않다. 어떤 시행인지 정해지면 표본 공간이 결정되고, 각 사건들에게 부여한 실현 가능성의 정도에 따라서 다양한 결과가 나타난다.

$\Omega = \{H, T\}$

안녕하세요. 홍박사입니다. 정말 오랜만에 포스팅을 합니다. 6개월만입니다. 핑계지만 연구 프로젝트 진행하느라 블로그에 거의 시간을 쏟지 못했습니다. 앞으로 분발하겠습니다. 지난 포스팅에서는 확률의 기본적인 정의에 대해서 알아 보았습니다. 이번 포스팅에서는 조금 더 나아가 "**확률 변수**"와 "**확률 분포**" 그리고 "**확률 함수**"에 대해서 이야기해 보겠습니다.

실수

'데이터 관점'에서 '확률변수'는 데이터의 'feature'임.

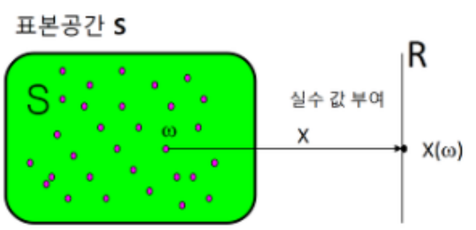
우선 변수(Variable)이란 무엇일까요? 변수란 특정 조건에 따라 변하는 값을 의미합니다. 그렇다면 확률변수(Random variable)는 무엇에 따라 변하는 값일까요? 당연히 확률에 따라 변하는 값이겠지요. **확률 변수**란 **무작위 실험을 했을 때, 특정 확률로 발생하는 각각의 결과를 수치적 값으로 표현하는 변수**를 말합니다. 한번에 이해하기에는 조금 어렵습니다. 예들 들어서 다시 설명해 보겠습니다. **확률 변수**는 **임의(Random)로 진행되는 실험(예: 동전을 무작위로 두번 던져서 그림 또는 숫자가 나오는 실험)에서 일정한 확률(예: 그림이 나올 확률 1/2, 그리고 뒤가 나올 확률 1/2)을 가지고 발생하는 결과에 실수 값(예: 앞=1, 뒤=0)을 부여하는 변수(variable)**를 말합니다.

↑ 즉, '확률변수'는 '함수'이다.

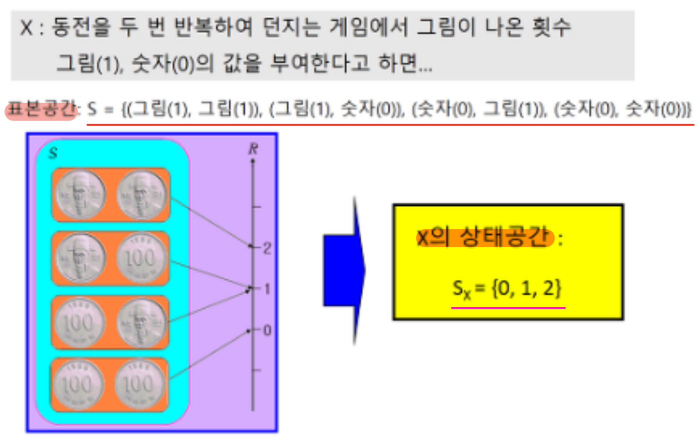
표본 공간을 대문자 "S", 그에 상응하는 실수 값을 부여하는 값을 "X", 그리고 부여된 실수 값 X에 따라 계산된 실수 값을 "R"이라고 했을때, 그 관계를 표현하면 아래 그림(왼쪽)과 같습니다. 이를 그림(오른쪽)과 같이 앞서 설명한 동전을 두번 반복해서 던지는 게임으로 설명하면 이해가 쉬울 것 입니다. 이 게임에서 표본공간(S)은 그림 또는 숫자가 나오는 조합이 되고, 각 변수에 그림 =1 숫자 = 0을 부여했을 때 (X) 나오는 값(R)은 {0,1,2}가 됩니다. 이를 확률 변수라고 합니다. 그리고 확률 변수(R)가 취하는 모든 실수들의 집합을 상태공간(State space)라고 하고, 그 상태공간를 구성하는 각 값이 나올 수 있는 가능성은 특정 확률(0=1/4, 1=1/2, 2=1/4의 확률)로 주어지게 됩니다.

상태공간

*표본 공간, 사건, 확률의 정의는 이전 포스트에서 확인하실 수 있습니다.
[통계 노트/통계 개념 정리] - [개념 통계] 표본공간, 사건 그리고 확률



상태공간(state space)
확률변수 x가 취하는 **모든 실수들의 집합**



확률 변수

확률적 데이터를 수학기호로 표시할 때는 변수를 표시할 때처럼 문자로 표시한다. 하지만 일반적인 변수가 특정한 하나의 숫자를 대표하는 변수인 것과 달리 확률적 데이터를 대표하는 변수는 나올 수 있는 값이 확률적 분포를 가진다. 즉 특정한 값은 자주 나오고 다른 어떤 값은 드물게 나올 수 있다. 이러한 변수를 확률 변수(random variable)라고 한다.

확률 변수는 숫자 혹은 벡터를 생성하는 기계에 비유할 수 있다.

확률 변수의 수학적 정의

수학적으로 확률 변수는 표본 공간을 정의역(domain)으로 가지고 실수를 공역(range)으로 가지는 함수로 정의한다. 즉 확률 변수는 표본 공간의 모든 ~~표본~~에 대해 어떤 실수 값을 붙인(할당한) 것이다.

표본점

X, Y 등의 대문자 알파벳을 사용하여 확률 변수를 표기하고 확률 변수에 의해 할당된 실수는 x, y와 같이 소문자 알파벳으로 표시하는 것이 보통이다. 경우에 따라서는 소문자 알파벳으로 확률 변수를 표기할 수도 있다.

$$\omega \in \Omega \ \xrightarrow{X} \ x \in \mathbf{R}$$

$$X(\omega) = x \ \ (x \in \mathbf{R})$$

확률과 확률 변수의 차이점은 다음과 같다.

- 확률은 표본으로 이루어진 집합 즉, 사건에 대해 할당된 숫자이지만 확률 변수는 ~~표본~~ 표본점 하나 하나에 대해 할당된 숫자이다.
- 확률은 0부터 1사이의 숫자만 할당할 수 있지만 확률 변수는 모든 실수 범위의 숫자를 할당할 수 있다.

확률 변수를 사용하면 모든 표본은 하나의 실수 숫자로 변하기 때문에 표본 공간을 실수의 집합 즉 **수직선 (number line)**으로 표시할 수 있다. 표본의 집합인 사건(event)은 이 수직선 상의 숫자의 집합인 **구간(interval)**으로 표시된다.

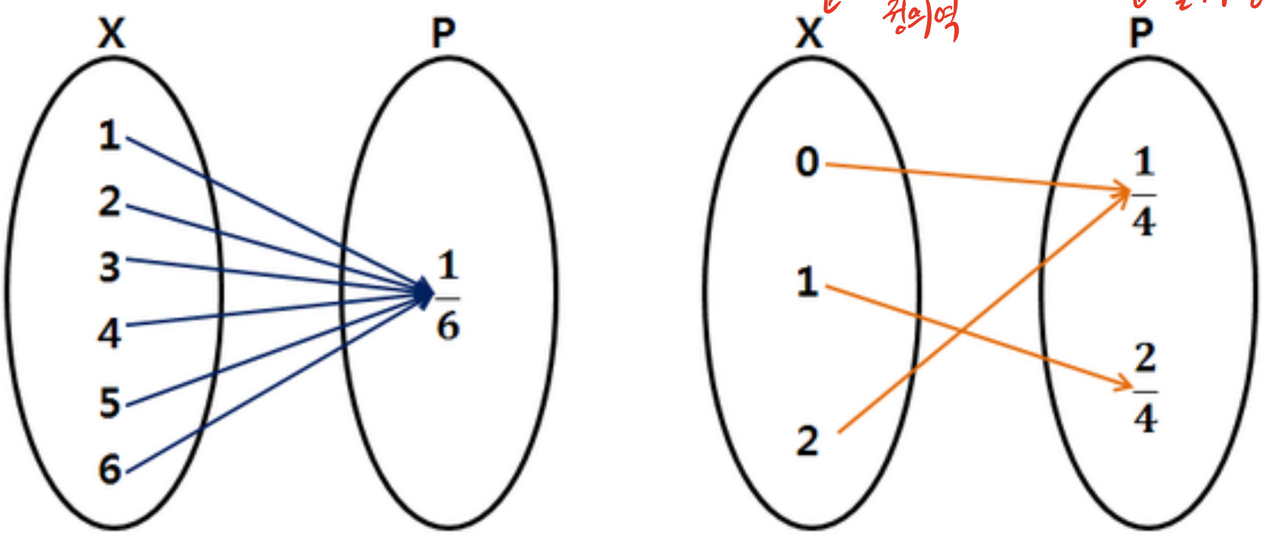
데이터 관점에서, '확률변수'는 데이터의 'feature'임.

1) 확률변수와 확률분포

확률변수란, 변수는 변수이되 그 값에 **확률의 개념이 수반된 변수**입니다. 이를테면 주사위를 던졌을 때 나오는 눈(1, 2, 3, 4, 5, 6)이나 동전 두 개를 던졌을 때 앞의 눈이 나오는 개수(0개, 1개, 2개)등이 확률변수가 될 수 있습니다. 각각의 확률변수의 값에 부여되는 확률은 서로 동일할 수도 있고(전자의 경우 1/6으로 모두 동일), 다를 수도 있습니다(후자의 경우 1/4, 2/4, 1/4로 서로 다름). 어찌됐건 확률변수가 정해지면 **그에 대응하는 확률**이 정해집니다. 이러한 대응관계, 즉, 하나의 변수에 대해 각각 확률이 하나씩 대응되는 함수 관계를 확률분포라고 명명합니다. 쉽게말해서, 확률분포는 확률변수에 해당하는 함수값, 즉 확률을 대응시킬 수 있는 **함수관계**입니다.

확률 변수의 치역이 정의역

0과 1사이의 실수가 공역.



위 그림에서 오른쪽은 앞의 예에서 확률변수가 주사위의 눈이 될 때의 확률분포를, 왼쪽은 확률변수가 두 동전을 던졌을 때 나오는 앞면의 개수일 때의 확률분포입니다. 이를 일반화시켜서 **표**로 만들면 아래와 같은 이산확률분포를 얻을 수 있습니다.