

2장. 확률밀도함수는 확률이 아니다?

◆ Introduction

생명보험회사에서는 보험료를 책정하기 위해 생명표를 이용한다. 보험회사의 생명표는 보험에 가입한 사람들을 대상으로 생존, 사망현상을 관찰하여 작성한 통계표를 의미하며 연령별 사망률, 기대수명, 잔여수명 등을 나타낸다. 아래의 표는 대한민국 전 국민을 대상으로 하여 통계청에서 발표한 2019년 생명표의 일부이다.

연령별	2019				
	기대여명(전체) (년)	사망확률(전체)	생존자(전체)	총생존년수(전체)	사망자(전체)
70세	17.1	0.01107	87,798	1,505,292	972
71세	16.3	0.01223	86,826	1,417,980	1,062
72세	15.5	0.01362	85,763	1,331,686	1,168
73세	14.7	0.01535	84,596	1,246,506	1,299
74세	14.0	0.01758	83,297	1,162,560	1,464
75세	13.2	0.02017	81,833	1,079,995	1,651

표 1) 2019년 국민생명표

출처: 통계청

이 표를 보고 2019년에 70세인 사람이 1년 이내 사망할 확률을 묻는다면 0.01107이라고 답할 것이다. 하지만 동일한 대상이 정확히 70.5세에 사망할 확률을 묻는다면 대답하기 힘들뿐더러 시간은 연속적이기 때문에 그 시점의 확률을 구하는 것도 쉽지 않다. 그리하여 생명표 같은 경우는 연령 별 구간의 사망률을 표기한다. 이번 장에는 점에서 확률의 의미를 갖진 않지만 구간에서 확률의 의미를 갖는 확률밀도함수(Probability density function: p.d.f.)와 확률변수(Random Variable)에 대해 알아보도록 하겠다. 확률밀도함수를 이해하려면 확률변수가 무엇인지를 알아야한다.

보험은 '확률'로부터 비롯되었다고 해도 과언이 아니다. 특정 현상에 '확률'이 있으면, 그 현상은 확실하지 않고 변동성이 있다 라는 말로 해석할 수 있다. 불안정함은 즉, 리스크를 뜻한다. 사람들은 리스크를 줄이기 위해 '대수의 법칙'을 이용하여 변동성이 있는 통계적 확률을 수학적 확률에 수렴시켰고, 이것이 바로 보험의 기본 원리이다.

확률변수는 특정 실험¹⁾으로부터 발생 가능한 모든 결과에 실수값(x)을 할당하는 함수 X 를 말한다. 그리고 확률변수에 할당된 모든 값에 대해 일어날 가능성을 나타낸 것이 확률분포이다. 확률변수는 이산적인 값을 대상으로 하는 이산확률변수와 연속적인 구간 내에서 임의의 실수값을 가지는 연속확률변수가 있다. 확률변수에 할당된 값들에 따라 취할 수 있는 확률분포가 다르다.

1) 실험은 특정 조건하에 현상을 관찰하는 것을 의미한다. 예를 들면 1억 보장 한도라는 조건하에 보험금 지급 현상을 관찰하면 실험에서 발생가능한 모든 보험금 지급 결과 값은 1억 이하로 한정된다.

2장. 확률밀도함수는 확률이 아니다?

먼저 이산확률변수에 대해 살펴보자. 이산확률변수 X 에 대한 확률분포함수는 확률질량함수로 표현한다. 확률질량함수는 이산확률변수 X 에 대해 $X=x$ 일 사건이 일어날 확률을 대응하는 함수로 정의한다. 예를 들어, 주사위의 눈을 확률변수 X 로 하는 사건을 시행할 때 $X=1$ 일 확률은 $1/6$ 이므로 $P(X=1)=1/6$ 이고 확률질량함수 $f(x)$ 는 사건 X 가 발생할 확률을 뜻한다.

반면 연속확률변수 X 에 대해서는 확률분포함수를 확률밀도함수로 정의한다. X 의 확률밀도함수는 정의상 다음과 같다.

$$f(x) = \frac{d}{dx}F(x) = -\frac{d}{dx}S(x)$$

즉, 확률밀도함수는 누적분포함수(Cumulative Distribution Function: c.d.f.)의 도함수로 정의한다. ($S(x)=1-F(x)$, $S(x)$ 는 생존함수이다.) 확률질량함수와 다르게 확률밀도함수는 누적분포함수의 도함수로 정의하는 이유에 대해 생각해보자. 예를 들어, 사람이 컵에 물을 따르는 시행을 한다고 하면 따른 물의 양은 연속확률변수 X 이다. 과연 사람이 컵에 정확하게 100ml를 따르는 확률은 몇일까? 이와 같은 시행에서 $X=100$ 인 확률은 0이다. 즉, 연속확률변수에서는 이산확률변수와 다르게 특정한 점에서의 확률을 정의할 수 없기 때문에 누적분포함수의 도함수로 정의한다.

따라서 확률밀도함수 $f(x)$ 는 확률이 아니다. 하지만 특정한 점이 아닌 구간에서는 확률의 의미를 갖는다. 컵에 정확하게 100ml의 물을 따를 확률은 0이지만 90ml~100ml를 따를 확률은 0이 아니다. 이처럼 확률밀도함수는 그 자체로 확률을 의미하지 않고 그 값 주변의 값 ($x \pm dx$)을 취할 가능성을 상대적인 높이로 나타낸다.

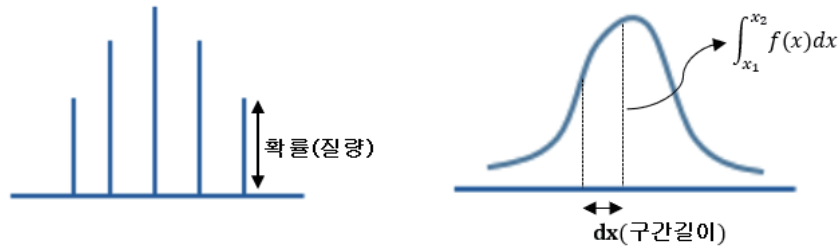
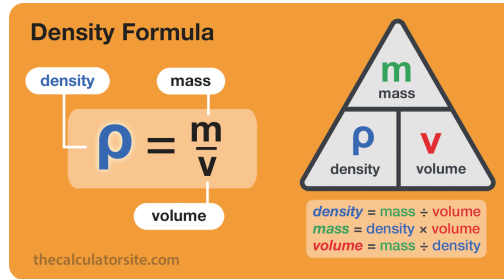
확률밀도함수의 수식적 정의로부터 다음이 성립한다.

$$F(x) = \int_0^x f(y)dy \quad S(x) = \int_x^{\infty} f(y)dy$$

따라서 다음이 항상 성립한다.

$$\int_0^{\infty} f(x)dx = 1$$

★ 쉬어가는 페이지 ★



“왜 확률질량함수, 확률밀도함수라고 부를까?”

확률을 가능성의 크기로 생각하여 질량으로 간주하면, 이산확률변수 X 값에 따라 확률(질량)을 나타낼 수 있으므로 확률질량함수라고 부른다. 물리학에서 밀도 = $\frac{\text{질량}}{\text{부피}}$ 이다. 통계학에서는 확률밀도함수를 $f(x) = \frac{\text{확률(질량)}}{\text{구간의길이(부피)}} = \frac{f(x)dx}{dx}$ 로 파악하고, 이는 밀도의 개념과 비슷하다.

확률밀도함수에서 구간 $[x_1, x_2]$ 의 확률은 $P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx$ 인데, 확률 $f(x)dx$ 를 구간 $[x_1, x_2]$ 에서 연속적으로 더해준다는 의미이다. 이 때 dx 는 x 의 변화량이므로 구간의 길이를 뜻한다. $f(x)$ 가 $\frac{\text{확률(질량)}}{\text{구간의길이(부피)}}$ 이면 $f(x)dx$ 가 확률의 의미를 갖는다는 것을 알 수 있다.

$$\frac{\text{확률}}{\text{구간의길이}} \times \text{구간의길이} = \text{확률}$$

$$\frac{f(x)dx}{dx} \times dx = f(x)dx$$

구간의 길이를 부피로 생각하면 확률을 질량으로 생각한다고 했으므로 $f(x)$ 는 $\frac{\text{질량(확률)}}{\text{부피(구간의길이)}}$ 가 되어 밀도를 의미한다. 따라서 연속확률변수 X 에 대한 $f(x)$ 를 확률밀도함수라고 부른다.

