Chapter 4. 연속형 확률변수

확률변수 X 가 가질 수 있는(서로 다른) 값이 유한(finite)이거나 셀 수 있는(countable) 이산형(discrete) 확률변수에 대한 살펴보았다. 여기서는 확률실험 결과 서로 다른 값이 무한히 발생할 수 있는 연속형(continuous) 확률변수에 대해 다룰 것이다. 연속형에서는 아무리 작은 구간을 설정하더라도 적어도 하나 이상의 값이 관측된다. (예) 키, 강우량, 소득, 수명, 경제 지표 등

이산형인 경우 확률밀도함수(확률변수가 갖는 값에 대해 확률을 할당하는 식, 표, 그래프)를 얻는 것은 어렵지 않다. 그러나 아무리 작은 구간이더라도 많은 값들이 관측될 수 있는 연속형의 경우 확률밀도함수를 얻는 것은 불가능하다. 그럼? 관측된 데이터로부터 히스토그램(histogram)을 그리고, 이것이 알려진 확률밀도함수(Gaussian 분포, t-분포) 중 어느 것과 가장 유사한지 판단하여 이론적 분포를 얻게 된다.

왜 우리는 확률밀도함수에 관심을 갖는가?

(1)표본 데이터의 확률밀도함수는 모집단의 확률밀도함수와 동일하다. 그러나 히스토그램에 의존하는 방식으로는 정확한 분포를 아는 것은 불가능하다.

(2)통계량 $h(X_1, X_1, ..., X_n)$ 의 확률밀도함수는 모집단의 모수(parameter)에 대한 추론에 핵심적인 역할을 한다. 표본의 분포를 알려면(물론 표본 평균은 CLT가 있기는 하지만) 모집단의 분포를 알아야 하는데 일반적으로는 불가능하다. 그러므로 데이터의 특성에 따라 모집단의 분포를 가정하게 된다.

4.1 분포함수 (distribution function)

정의(DEFINITION)

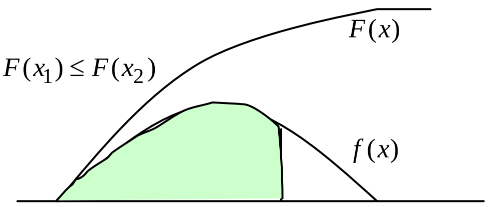
연속형 확률변수 X 의 확률분포함수(probability distribution function) 혹은 누적확률밀도함수(cumulative density function) $F(x)$ (기호)는 $F(x) = P(X \le x),\ for\ -\infty < x < \infty$ 라 정의한다.

$F(x)$ 의 성질

(1) $F(-\infty) = 0, F(\infty) = 1$

(2) $F(x)$ 는 비감소(non-decreasing) 함수이다. $If\ x_1 < x_2$, then $F(x_1) \le F(x_2)$

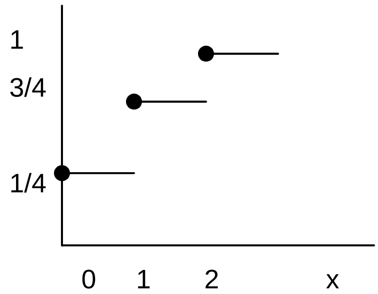
초록색 부분이 $F(x)$ 이다.



EXAMPLE 4-1 분포함수 구하기

이산형 확률변수 $X \sim Binomial(n=2, p=0.5)$ 에 대하여 분포함수 $F(x)$ 을 얻으시오.

$$F(x)=\begin{cases} 0, x<0 \\ 1/4, 0 \le x < 1 \\ 3/4, 1 \le x < 2 \\ 1, x \ge 2 \end{cases}$$ (이산형 확률변수의 $F(x)$ 는 step 함수)



정의(DEFINITION)

(1)만약 $F(x)$ 가 연속이면 확률변수 X 는 연속형이다.

(2)연속형 확률변수의 분포함수를 $F(x)$ 라 하면 확률밀도함수 $f(x)$ 는 다음에 의해 계산된다. $f(x)=\frac{d}{dx}F(x)$

$f(x)$ 는 $F(x)$ 으로부터 얻어지므로 아래 성질을 갖는다.

연속형 확률밀도함수의 성질

(1) $f(x) \ge 0$ for any value of x

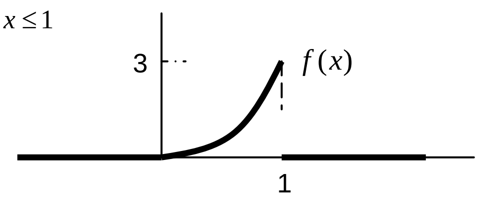
(2) $\int_{-\infty}^{\infty} f(x)dx = 1$

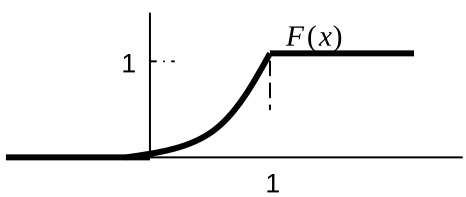
(3) $P(a \le X \le b) = F(b) - F(a)$

EXAMPLE 4-2 분포함수 구하기(2)

확률변수 X 의 확률밀도함수 $f(x)=3x^2, 0 \le x \le 1$ 일 경우 분포함수 $F(x)$ 을 구하고 그래프를 그리시오.

$$F(x)=\begin{cases} 0, x<0 \\ x^3, 0 \le x \le 1 \\ 1, x \ge 1 \end{cases}$$





EXAMPLE 4-3 분포함수 이용하기

확률변수 X 의 확률밀도함수가 $f(x) = cx^2, 0 \le x \le 2$ 일 경우

(1)상수 c 을 구하시오. (2)분포함수 $F(x)$ 구하시오. (3) $P(1 < X \le 2)$ 을 계산하시오.

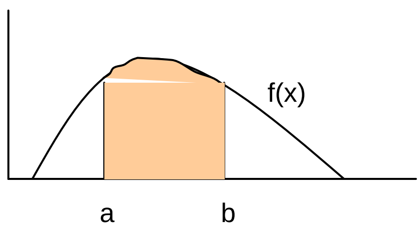
$c = 3/8$

$F(x) = x^3/8$, $0 \le x \le 2$

$7/8$

확률과 분포함수

$$P(a \le X \le b) = F(b) - F(a)$$



HOMEWORK #9-1 DUE 4월 14일(목)

확률변수 X 의 확률밀도함수가 $f(x) = cx, 0 \le x \le 2$ 일 경우.

(1)상수 c 을 구하시오.

(2)확률분포함수 $F(x)$ 을 구하시오.

(3) $F(x)$ 이용하여 $P(1 < X \le 2)$ 계산하시오.

HOMEWORK #9-2 DUE 4월 14일(목)

확률변수 X 의 분포함수는 $F(x) = \begin{cases} 0, x \le 0 \\ x/8, 0 < x < 2 \\ x^2/16, 2 \le x < 4 \\ 1, x \ge 4 \end{cases}$ 이다.

(1)확률밀도함수 $f(x)$ 을 구하시오.

(2) $P(X > 1.5)$ 계산하시오.

(3) $P(X \ge 1 | X \le 3)$ 계산하시오.

4.2 기대값(Expected value)

□평균 mean($\mu, \bar{x}$): 데이터의 중앙 위치 모수 $E(X)$

□분산 variance(σ^2, s^2): 산포(흩어짐) 모수 $E(X-E(X))^2 = E(X^2) - E(X)^2$

□표준편차 standard deviation(σ, s): 분산의 제곱근

연속형 확률변수의 확률밀도함수(pdf)를 얻는 것은 불가능하지만, 적률($E(X^k)$)과 경험적 법칙(Empirical Rule, $\pm 2\sigma$, 95%)이나 Tchebysheff's inequality($P(|X-\mu| < k\sigma) \ge 1 - \frac{1}{k^2}$)을 이용하여 확률변수(데이터)의 분포를 판단할 수 있다.

정의(DEFINITION)

□연속형 확률변수 X 의 기대값(expected value)은 $E(X) = \int x f(x) dx$ 으로 정의한다.

□확률변수 X 의 함수 $g(X)$ 의 기대값은 $E(g(X)) = \int g(x) f(x) dx$ 로 정의된다.

□만약 $g(X) = (X - E(X))^2$ 이면, $g(X)$ 의 기대치는 확률변수 X 의 분산이다.

정리(THEOREM)

(1)상수 c 에 대하여 $E(c) = c$ 이 성립한다.

$(2) E[cg(X)] = cE[g(X)]$

$(3) E[g_1(X) + g_2(X) + ... + g_k(X)] = E[g_1(X)] + E]g_2(X)] + ... + E]g_k(X)]$

PROOF obvious

HOMEWORK #9-3 DUE 4월 14일(목)

연속형 확률변수 X 는 평균 μ 이고 분산 σ^2 을 갖는다. 상수 a,b 에 대하여 다음이 성립함을 보이시오. $E(aX+b) = a\mu + b$ and $V(aX+b) = a^2\sigma^2$

EXAMPLE 4-4 평균과 분산 구하기

만약 $X \sim f(x) = 1/2, 59 \le x \le 61$ 이면, 확률변수 X 의 평균과 분산을 구하시오.

60 / 0.333

EXAMPLE 4-5 평균과 분산 구하기

만약 $X \sim f(x) = (3/2)x^2, 0 \le x \le 1$ 이면, 확률변수 $W = (5 - 0.5X)$ 의 평균과 분산을 구하시오.

4.8 / 0.039

HOMEWORK #9-4 DUE 4월 14일(목)

확률밀도함수 $f(x) = 2x, 0 \le x \le 1$ 을 갖는 확률변수 X 에 대하여.

(1) $E(X), V(X)$ 을 계산하시오.

(2)확률변수 $W = 200X - 60$ 의 평균과 분산을 구하시오..

4.3 균일분포 (Uniform Dist.)

버스는 오전 8:00와 8:10 사이에 반드시 정차하고 10분 사이 어느 구간에서든 버스가 도착하는 가능성은 동일하다고 하자. 즉 임의의 구간에 버스가 도착할 확률은 시간 길이에 비례한다.

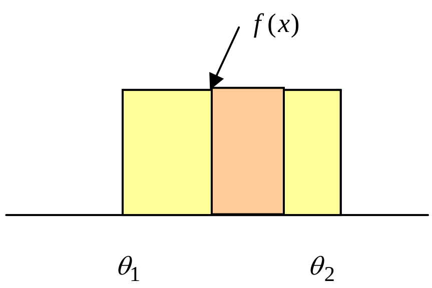
정의(DEFINITION)

확률변수 X 을 임의의 구간(θ_1, θ_2) 사이의 시간이나 거리라 하자. 만약 X 의 pdf가 다음과 같다면 확률변수 X 는 구간 (θ_1, θ_2) 에서 연속형 균일분포를 갖는다고 한다.

$$f(x) = \frac{1}{\theta_2 - \theta_1}, \theta_1 \le x \le \theta_2$$

NOTATION $X \sim Uniform(\theta_1, \theta_2)$

In SAS: PDF('UNIFORM', x, θ_1, θ_2)



$\theta_1 = 0, \theta_2 = 1$ 인 균일분포는 난수(random number) 생성에 이용된다.

평균과 분산

$$E(X)=\frac{\theta_1+\theta_2}{2},\quad V(X)=\frac{(\theta_2-\theta_1)^2}{12}\ (\text{숙제})$$

EXAMPLE 4-6 균일분포 확률

임의의 30분 동안 은행에 고객이 방문하는 시간은 균일 분포를 따른다고 하자. 마지막 5분에 고객이 도착할 확률을 계산하시오.

1/6

EXAMPLE 4-7 균일분포 확률(2)

확률변수 $X \sim Uniform(50,70)$ 라면 $P(X \ge 65 \mid X \ge 55)$ 을 구하시오.

1/3

HOMEWORK #10-1 DUE 4월 19일(화)

$X \sim Uniform(\theta_1, \theta_2)$ 의 평균과 분산이 각각 $E(X)=\frac{\theta_1+\theta_2}{2}$, $V(X)=\frac{(\theta_2-\theta_1)^2}{12}$ 임을 보이시오.

HOMEWORK #10-2 DUE 4월 19일(화)

만약 낙하산이 marker A와 B 사이에 임의의 지점에 떨어진다고 하자.

(1)낙하산이 B 보다 A 지점에 더 가까이 떨어질 확률을 구하시오.

(2)낙하산 떨어진 지점에서 지점 A까지의 거리가 지점 B까지의 거리의 3배 이상일 확률을 구하시오.

(3)3개의 낙하산 중 정확하게 한 개만 지점 B에 가까이 떨어질 확률을 구하시오.