1.

2. Probability: Univariate Models

2.1. Introduction

2.1.1. What is probability?

통계에서 확률을 바라보는 관점에는 2가지가 존재함 frequentist , bayesian
frequentist : 확률을 장기적으로 일어나는 사건의 빈도로 보는 것. 빈도주의
bayesian : 확률을 사건 발생에 대한 믿음 또는 척도로 바라보는 관점. 베이지안
베이지안 관점의 장점은 일회성 사건에 대한 불확실성도 계산할 수 있음. 이 책에서는 베이지안 관점으로 봄.

빈도주의와 베이지안 관점 차이 예시

a. 동전을 던졌을 때 앞/뒷면이 나오는 사건의 확률
    i. 빈도주의 : 동전을 던져 앞면이 나오는 사건의 '확률'은 0.5이다.
    ii. 베이지안 : '앞면이 나왔다'는 주장의 신뢰도가 0.5이다.
b. 검진 결과에 의해 암에 걸렸을 확률이 90%이다.
    i. 빈도주의 : 이러한 검진 결과를 가진 환자는 정밀검사를 했을때 100에 90명은 암에 걸렸다.
    ii. 베이지안 : 자신이 암에 걸렸음을 주장하는 의사의 주장이 (신뢰도) 사실일 가능성이 90%이다.

2.1.2. Types of uncertainty

epistemic uncertainty : 주어진 데이터 세트를 가장 잘 설명하는 최상의 모델 매개변수 및 모델 구조의 불확실성 . 예를들어 데이터셋에 대한 세가지 모델중 어떤 모델이 가장 적합한 모델이 되는 지에 대한 불확실성 (**model uncertainty**)

aleatoric uncertainty : 학습 데이터 자체에 노이즈가 많아져서 불확실성이 생기는 경우. 예를 들어 학습할때 분류할 개,고양이,소가 있다고 할때 고양이 이미지만 노이즈가 있고 개와 소 이미지는 정상적인 이미지인 경우에 발생하는 불확실성 (**data uncertainty**)

2.1.3. Probability as an extension of logic

2.1.3.1. Probability of an event

Pr(A) : 사건 A가 참이라고 믿을 확률 (A가 발생할 장기적 비율) .
$0 \leq Pr(A) \leq 1$.
Pr(A) = 0 : 사건 A가 발생하지 않음.
Pr(A) = 1 : 사건 A가 발생함
$Pr(\bar{A}) = 1 - Pr(A)$. 사건 A가 일어나지 않을 확률

2.1.3.2. Probability of a conjunction of two events

joint probability : 조건부확률.

두개의 서로 다른 사건이 동시에 일어날 확률 : $Pr(A \wedge B) = Pr(A,B)$
A와 B가 독립 사건이면 : $Pr(A,B) = Pr(A)Pr(B)$

2.1.3.3. Probability of a union of two events

$Pr(A \vee B) = Pr(A) + Pr(B) - Pr(A \wedge B)$
사건 A,B가 상호배타적이라 동시에 발생할 수 없으면 $Pr(A \wedge B) = 0$.

2.1.3.4. Conditional probability of one event given another

A가 발생했다는 가정하에 B가 일어날 조건부 확률

$$\Pr(B|A) \triangleq \frac{\Pr(A, B)}{\Pr(A)}$$

($Pr(A) = 0$ 인 경우는 정의하지 않음. 사건 A는 발생했다고 가정했음)

2.1.3.5. Independence of events

사건 A와 사건 B가 독립 : $Pr(A, B) = Pr(A)\ Pr(B)$

2.1.3.6. Conditional independence of events

사건 A와 B는 사건 C에 대해 조건부 독립 : $Pr(A, B|C) = Pr(A|C)\ Pr(B|C)$

2.2. Random variables

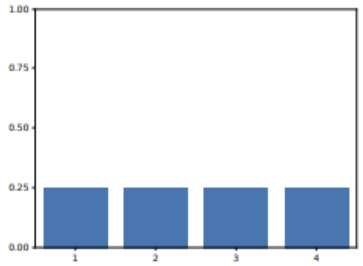
확률 변수는 사건에 실수값을 대응시키고 그 값에 확률을 부여한 것.
확률 공간의 점들의 집합은 표본 공간 sample space라고 한다.

확률 함수는 확률을 가진 어떤 사건이 일어날 확률을 통해 파라미터를 만들고 이를 활용해 함수를 만드는 것. (확률 p를 가진 어떤 사건이 n회 시행 중에서 x회 나타날때, 확률 변수 x 와 이에 대응되는 p(x)의 관계를 나타낸 함수)

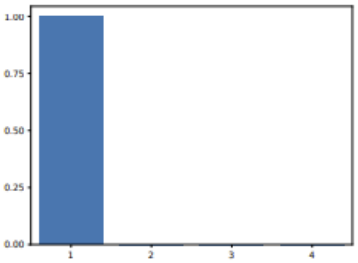
2.2.1. Discrete random variables

Discrete random variables : 이산 확률 변수 X.
X는 유한집합이나 가산무한 집합으로부터 어떠한 값도 가질 수 있다. 셀 수 있는 특정 값으로 구성되거나 일정 범위로 나타나는 경우를 뜻함. 예를 들어 주사위 3번 던질 때 1이 몇번 나오는가.
X = x 의 확률은 p(X=x)로 나타내고 간단히 p(x)라고 한다.

probability mass function : 확률 질량 함수. pmf.
이산확률 변수를 나타내는 함수. 0과 1 사이의 값을 가지며 총합은 1이 된다는 조건을 만족한다. 유한한 상태 공간에서 정의된 pdf 이며 왼쪽은 균등 분포, 오른쪽은 퇴화 분포이다. (퇴화 분포 : 이산 확률 변수가 하나의 값을 가질 확률이 1인 분포. X는 항상 1과 같다는 것을 표현함.)



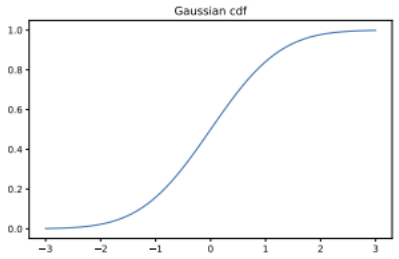
(a)



(b)

2.2.2. Continuous random variables

Continuous random variables : 연속 확률 변수는 변수가 연속적인 숫자이거나 무한한 경우와 같이 셀 수 없는 경우를 뜻한다. 예를 들어 각 반별 학생의 평균 키. P(x) = Pr(X ≤ x)



Cumulative distribution function : 누적 분포 함수. cdf.
A = (X ≤ a), B = (X ≤ b) and C = (a < X ≤ b),a < b .
a 와 b는 상호 배타적이므로 Pr(B) = Pr(A) + Pr(C).

P(x) = Pr(X ≤ x) 누적분포함수 정의

Pr(a < X ≤ b) = P(b) − P(a) : x를 불확실한 연속량으로 가정한다면 x가 a이상 b 이하일 확률은 다음과 같이 계산한다. 임의의 구간에 있을 확률

Probability density function : 확률 밀도 함수. pdf.
연속적인 변수에 의한 확률 분포 함수.

$$p(x) \triangleq \frac{d}{dx}P(x)$$

$$\Pr(a < X \leq b) = \int_a^b p(x)dx = P(b) - P(a)$$ 연속 변수가 유한 구간에 있을 확률

$$\Pr(x < X \leq x + dx) \approx p(x)dx$$ 간격이 좁아지면 이렇게 표현 가능
직관적으로 X가 x주위 구간에 있을 확률이 구간 너비의 x배에 해당하는 밀도 라는 뜻

cdf 가 증가하기 때문에 역함수를 가짐. cdf의 역함수는 ppf(percent point function)임. p 가 x의 cdf인 경우 $p^{-1}(q)$ 는 Pr(X ≤ xq) = q를 만족하고 이걸 p의 q번째 분위수 라고 함. $p^{-1}(0.5)$는 중앙값, $p^{-1}(0.25)$, $p^{-1}(0.75)$는 사분위수 임.

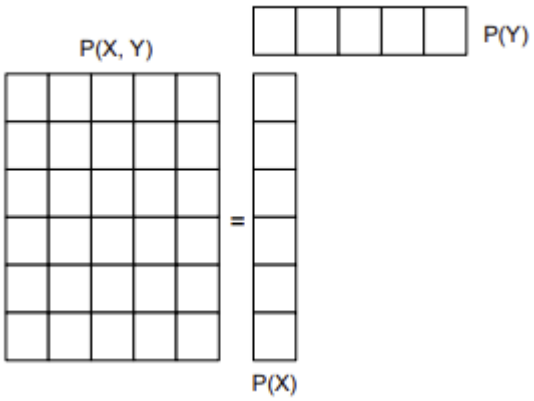
2.2.3. Sets of related random variables

먼저 두 개의 확률 변수 X와 Y가 있다고 가정.
X와 Y의 가능한 모든 값에 대해 p(x, y) = p(X = x, Y = y)를 사용하여 두 확률 변수의 공동 분포를 정의할 수 있음.

$p(X,Y)$	$Y=0$	$Y=1$
$X=0$	0.2	0.3
$X=1$	0.3	0.2

두 변수 모두 유한 cardinality를 갖는 경우 결합 분포를 모든 항목의 합이 1이 되는 표로 나타낼 수 있다.

두 변수가 독립이면 두 변수를 곱하고, 두 변수가 finite cardinality를 가지면 2d 표를 1d 표의 곱으로 인수분해할 수 있음 .



결합 분포가 주어지면 확률 변수의 한계 분포를 이렇게 정의함 >>>> y의 가능한 상태를 다 더함

$$p(X = x) = \sum_y p(X = x, Y = y)$$

확률 변수의 조건부 분포 정의 >>>

$$p(Y = y|X = x) = \frac{p(X = x, Y = y)}{p(X = x)}$$

이 방정식을 정리하면
product rule 이라고 불리는 *p(x, y) = p(x)p(y|x)* 가 됨
이 규칙을 D변수로 확장하면
p(x1:D) = p(x1)p(x2|x1)p(x3|x1, x2)p(x4|x1, x2, x3). . . p(xD|x1:D−1) 를 얻을 수 있음 (고차원 결합 분포 생성 방법 제공)

2.2.4. Independence and conditional independence

X와 Y의 결합 확률을 곱으로 표현할 수 있다면 무조건적인 독립 또는 marginally 독립이라고 하며 X⊥Y 로 표현한다.
일반적으로 결합 확률이 marginal의 곱이면 변수 집합이 상호 독립적이라고 한다.
X ⊥ Y ⇐⇒ p(X, Y) = p(X)p(Y) : X와 Y가 독립

$$p(X_1, \ldots, X_m) = \prod_{i=1}^{m} p(X_i)$$

대부분의 변수는 다른 변수에게 영향을 주는 경우가 많기 때문에 무조건적인 독립은 매우 드물다. 하지만 이런 영향은 직접적이기보다는 다른 변수에 의해 중재된다. 그러므로 조건부 결합확률이 조건부 주변의 곱으로 표현될 수 있는 경우에만 “Z가 주어졌을때 X와Y는 조건부 독립이다” 라고 한다.

2.2.5. Moments of a distribution

2.2.5.1. Mean of a distribution

분포를 설명할 때 가장 익숙한 속성은 평균 또는 기댓값이며, μ와 같이 표기한다.
연속확률 변수에서 평균은 다음과 같이 정의된다.

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x\, p(x) dx$$

이산 확률 변수에서 평균은 다음과 같이 정의된다.

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x\, p(x)$$

평균은 선형 연산자이므로 E [aX + b] = aE [X] + b
가 성립함.
이것을 **linearity of expectation** 기댓값의 선형성이라함.

2.2.5.2. Variance of a distribution

분산은 분포의 확산(흩어진 정도)을 측정한 것으로 σ^2로 표시하고 다음과 같이 정의함.

$$\mathbb{V}[X] \triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x) dx$$
$$= \int x^2 p(x) dx + \mu^2 \int p(x) dx - 2\mu \int x p(x) dx = \mathbb{E}\left[X^2\right] - \mu^2$$

표준편차는 다음과 같이 정의함

$$\mathrm{std}[X] \triangleq \sqrt{\mathbb{V}[X]} = \sigma$$

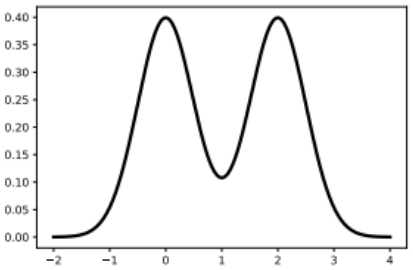
$V[aX + b] = a^2 V[X]$가 성립함.

2.2.5.3. Mode of a distribution

분포의 최빈값은 밀도나 질량이 가장 높은 값.

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmax}}\, p(\boldsymbol{x})$$

분포가 다중모드인경우 고유하지 않을 수 있다.



2.2.5.4. Conditional moments

두 개 이상의 종속 확률 변수가 있는 경우 하나가 다른 하나에 대해 알고 있는 순간을 계산할 수 있다. 예를 들어, 전체 기대의 법칙 law of total expectation

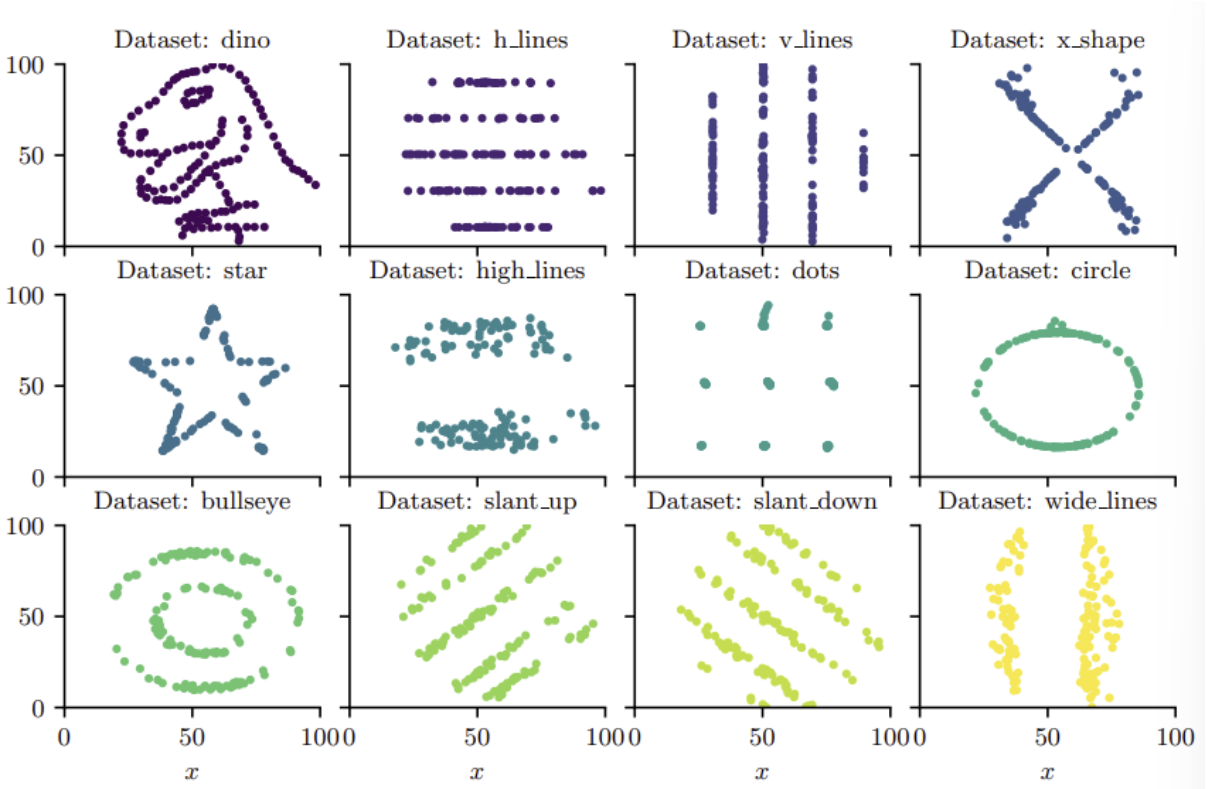
이라고도 불리는 반복 기대의 법칙 law of iterated expectations는
$E[X] = E_Y[E[X|Y]]$ 이다.

X,Y가 이산형 랜덤 변수라고 가정하면 이렇게 증명할 수 있다.

$$\mathbb{E}_Y[\mathbb{E}[X|Y]] = \mathbb{E}_Y\left[\sum_x x\, p(X=x|Y)\right]$$
$$= \sum_y \left[\sum_x x\, p(X=x|Y=y)\right] p(Y=y) = \sum_{x,y} xp(X=x, Y=y) = \mathbb{E}[X]$$

2.2.6. Limitations of summary statistics

확률 분포를 평균, 분산 등의 간단한 통계를 사용해 요약하는 것이 일반적이지만 이로인해 많은 정보가 손실될 수 있다.--> **Anscombe's quartet**에서 볼 수 있다.
(분산, 표준편차, 평균까지 같지만 실제 데이터는 완전 다른)



2.3. Bayes' rule

베이즈 정리는 조건이 주어졌을때의 조건부 확률을 구하는 공식이다.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

여기에서 P(A) 는 **prior distribution**, 사전 확률이라고 하며, 사건 B가 발생하기 전에 가지고 있던, 알고있던 사건 A의 확률이다.
책에서는 H=h가 A, Y =y가 B라고 보면 같은 꼴이다.

$$p(H = h|Y = y) = \frac{p(H = h)p(Y = y|H = h)}{p(Y = y)}$$

likelihood : 가능성. 관측된 사건이 고정된 상태에서 확률 분포가 변화될때 (확률 분포를 모를때, 가정할때) 확률을 표현하는 단어.(어떤 값이 관측되었을때, 해당 관측값이 어떤 확률 분포로부터 나왔는지에 대한 확률.. 고정되는 요소 = 관측값.)
사전 확률에 곱하는 P(B|A)가 바로 가능도이다.

posterior distribution : 사건 B가 발생하면 이 정보를 반영해 사건 A의 확률은 P(A|B) 라는 값으로 변하게 되며 이를 사후확률 이라고 함 사후 확률은 사전확률에 P(B|A)/P(B)라는 값을 곱하면 얻을 수 있다.

- P(A|B) : 사후확률(posterior). 사건 B가 발생한 후 갱신된 사건 A의 확률
- P(A) : 사전확률(prior). 사건 B가 발생하기 전에 가지고 있던 사건 A의 확률
- P(B|A) : 가능도(likelihood). 사건 A가 발생한 경우 사건 B의 확률

베이즈 정리는 사건 B가 발생함으로써 (= 사건 B가 진실이라는 것을 알게 됨으로써, 즉 사건 B의 확률 P(B) = 1이라는 것을 알게 됨으로써) 사건 A 의 확률이 어떻게 변화하는지를 표현한 정리. 즉 새로운 정보가 기존의 추론에 어떻게 영향을 미치는지를 나타낸다.

2.3.1. Example: Testing for COVID-19
2.3.2. Example: The Monty Hall problem
2.3.3. Inverse problems

2.4. Bernoulli and binomial distributions

2.4.1. Definition

Bernoulli distribution 베르누이 분포
결과가 두가지 중 하나로 나오는 실험, 시행을 베르누이라고 한다. 동전을 던져 앞,뒷면이 나오게 하는 것도 베르누이 시행이다.
베르누이 시행의 결과를 실수 0또는 1로 바꾼 것을 베르누이 확률 변수

$$\mathrm{Ber}(y|\theta) = \begin{cases} 1-\theta & \text{if } y = 0 \\ \theta & \text{if } y = 1 \end{cases}$$

binomial distribution 이항 분포
성공확률이 μ 인 베르누이 시행을 N번 반복하는 경우를 생각해보자. 가장 운이 좋을 땐 N번 모두 성공하고, 가장 운이 나쁜 경우엔 한번도 성공하지 못할

것이다. N번중 성공한 횟수를 확률변수 X라고 한다면 X의 값은 0부터 N까지의 정수중 하나가 될 것이다. 이런 확률 변수를 이항분포를 따르는 확률 변수라고 한다.

베르누이분포와 이항분포는 모두 베르누이 확률변수에서 나온 표본값이다. 표본 데이터가 하나 뿐이면 베르누이분포가 되고 표본 데이터가 여럿이면 이항분포가 된다.

2.4.2. Sigmoid (logistic) function

시그모이드란 머신러닝에서 주로 사용되는 S형 함수이다. 0과 1 사이의 값을 출력한다. 이 함수는 확률과 같은 비율을 나타내는데 사용된다.

시그모이드 함수는 입력값이 어떤 범위에 들어갈때 출력 값 0과 1 사이에서 급격하게 변하는 특성을 갖고 있다.

이진 변수를 예측하려면 다음과 같은 조건부 확률 분포를 사용해야한다.

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y|f(\boldsymbol{x}; \boldsymbol{\theta}))$$

$0 \le f(x; \theta) \le 1$를 피하기 위해 $p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y|\sigma(f(\boldsymbol{x}; \boldsymbol{\theta})))$ 이 모델을 사용한다.

시그모이드 함수는 다음 시그마와 같이 정의된다. $\sigma(a) \triangleq \frac{1}{1+e^{-a}}$

2.4.3. Binary logistic regression

로지스틱 회귀란 종속 변수가 범주형인 경우에 적용하는 회귀 분석 기법이다. 종속 변수가 두개의 범주 중 하나에 속할 확률을 예측한다.

2.5. Categorical and multinomial distributions

2.5.1. Definition

베르누이 확률변수는 0이나 1이 나오는 확률 변수였고 동전을 던져 나오는 결과를 묘사할 때 쓸 수 있다. 동전이 아닌 주사위를 던져서 나오는 경우는 어떻게 묘사할 수 있을까 ? = 카테고리 확률 변수

카테고리 확률 변수는 1부터 k까지 k개 정수값 중 하나가 나온다. 이 정숫값을 범주값, 카테고리라고 한다. 주사위를 던져 나오는 눈금의 수는 k = 6인 카테고리 분포이다.

원래 카테고리는 스칼라값이지만 카테고리 확률 변수는 1과 0으로 이뤄진 다차원 벡터를 출력한다. 숫자를 이렇게 변형하는 것을 One-Hot-Encoding 원핫인코딩이라고 한다.

카테고리 확률 변수의 확률 분포인 카테고리 확률 분포는 이렇게 표현한다.

$$\text{Cat}(\boldsymbol{y}|\boldsymbol{\theta}) \triangleq \prod_{c=1}^{C} \theta_c^{y_c}$$

원핫인코딩을 쓴 덕분에 이렇게 간단하게 표현할 수 있다.

2.5.2. Softmax function

softmax 함수는 조건부확률로 정의하면 아래와 같다.

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \text{Cat}(y|f(\boldsymbol{x}; \boldsymbol{\theta}))$$

소프트 맥스 함수는 다중 클래스 분류 모델을 만들 때 사용한다. 결과를 확률로 해석할 수 있게 변환해주는 함수로 높은 확률을 가지는 class로 분류한다. 이는 결과값을 정규화시키는 것으로도 생각할 수 있다.

소프트맥스 함수의 출력은 분류하고자하는 클래스의 개수만큼 차원을 가지는 벡터로 각 원소는 0과 1 사이의 값을 가지며, 이 각각은 특정 클래스가 정답일 확률을 나타낸다.

$$\text{softmax}(\boldsymbol{a}/T)_c = \begin{cases} 1.0 & \text{if } c = \text{argmax}_{c'} a_{c'} \\ 0.0 & \text{otherwise} \end{cases}$$

2.5.3. Multiclass logistic regression

다항 로지스틱 회귀 분석.
y의 범주가 3개 이상이며 명목형일때 사용하는 로지스틱 회귀 분석이다.

2.5.4. Log-sum-exp trick

2.6. Univariate Gaussian (normal) distribution

단변량 가우시안 분포

2.6.1. Cumulative distribution function

연속 확률변수 y의 누적 분포 함수 (cdf)를 다음과 같이 정의한다.

$$P(y) \triangleq \Pr(Y \leq y)$$

누적 확률 분포 함수는 확률 변수가 특정 값보다 작거나 같을 확률을 나타낸다. 이걸 사용하면 임의의 구간에 있을 확률도 구할 수 있다.

$$\Pr(a < Y \leq b) = P(b) - P(a)$$

2.6.2. Probability density function

확률 밀도 함수 pdf . (누적 분포 함수를 미분한 것)

연속형 확률 변수가 특정 구간에 속할 확률을 나타낼 수 있다.