8.7 베타분포, 감마분포, 디리클레분포

베타분포 / 감마분포 / 디리클레분포 → "모수를 조절하여 분포모양을 우리가 원하는대로 쉽게 바꿀수 있다"

★ 이러한 특성 때문에, 이 분포들은 데이터가 따르는 분포를 표현하기보다는 베이지안 확률론의 관점에서 어떤 값에 대해 우리가 가지고 있는 확신 혹은 신념의 정도를 표현하는데 주로 사용된다. (for 베이지안 추정)

베타분포 scipy.stats.beta(a,b)

베타분포 (Beta distribution)는 a와 b라는 두 모수를 가지며, 표본공간은 0과 1 사이의 실수다. 즉 0과 1 사이의 표본값만 가질 수 있다.

$$Beta(x; a, b), \quad 0 \le x \le 1$$

베타분포의 확률밀도함수는 다음과 같다.

$$Beta(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \underbrace{x^{a-1}(1-x)^{b-1}}_{\text{베르누이 분포 유사}}$$

이 식에서 $\Gamma(a)$는 감마함수 (Gamma function)라는 특수함수로 다음과 같이 정의된다.

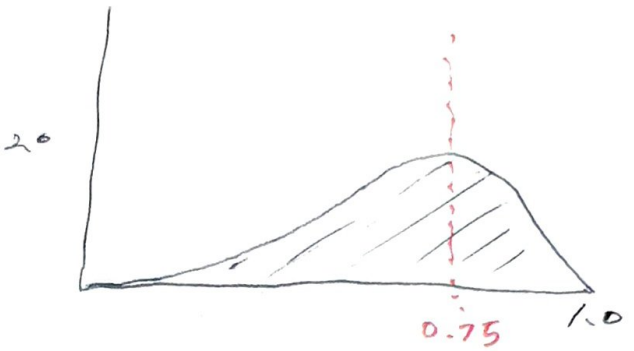
$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

★ 베타분포의 확률밀도함수를 비교적 간단한데, 사람이 인위적으로 만들어서 그렇다. 정규분포 및 통계쪽 분포 (t, chi2, F 등)는 자연계의 법칙을 수식화한 것이기 때문!

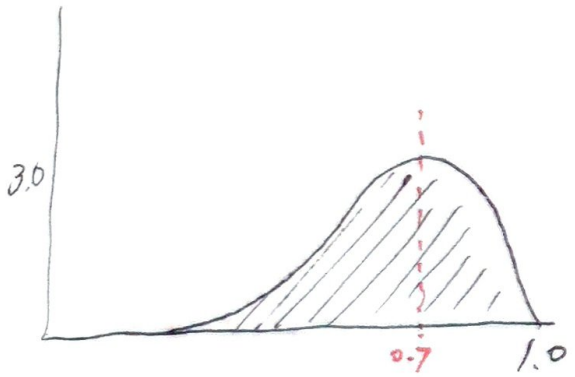
베타분포 모수별 pdf 모양.

(A) $a=1$, $b=1$

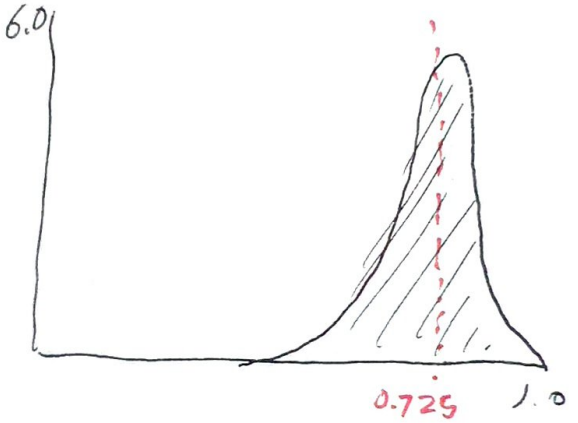
(B) $a=4$, $b=2$, 최빈값 $=0.75$



(C) $a=8$, $b=4$, 최빈값 $=0.7$



(D) $a=30$, $b=12$, 최빈값 $=0.725$



베타분포의 기댓값, 최빈값, 분산

기댓값 $E[X] = \dfrac{a}{a+b}$

최빈값 : 확률밀도가 가장 커지는 위치

$mode = \dfrac{a-1}{a+b-2}$ $\Rightarrow$ $a=b$일 때, $x=0.5$에서 가장 확률밀도가 커짐.

분산 : 확률분포의 폭

$Var[x] = \dfrac{ab}{(a+b)^2(a+b+1)}$ $\Rightarrow$ a, b가 커질수록 분산(확률분포의 폭)이 작아진다.

베타분포와 베이지안 추정

~~베타분포란~~

베타분포는 0부터 1까지의 값을 가질수 있는 베르누이분포의 모수 μ의 값을 베이지안 추정한 결과를 표현한 것이다.

- 베이지안 추정은 모수가 가질수 있는 모든 값에 대해 가능성을 확률분포로 나타낸 것을 말한다.
- 실제로 베르누이분포의 모수를 베이지안 추정하는 것은 내용이 다루게 된다. (여기서는 결과만 보임)

앞서 살펴본 4개의 베타분포 그림이 베이지안 추정 결과라면,
각각의 그림은 베르누이분포의 모수 μ에 대해 다음과 같이 추정한 것과 같다.

- (A) 베르누이분포의 모수 μ를 추정할 수 없다 (=정보없음)
- (B) 베르누이분포의 모수 μ값이 0.75일 가능성이 가장크다 (정확도 낮음)
- (C) 베르누이분포의 모수 μ값이 0.70일 가능성이 가장 크다 (정확도 중간)
- (D) 베르누이분포의 모수 μ값이 0.725일 가능성이 가장크다 (정확도 높음)

연습문제 8.7.1

베르누이 모수를 추정한 결과가 $\mu = \frac{1}{3}$이고, 추정오차(표준편차)가 0.2라고 하자.
이 추정결과를 나타내는 베타분포의 모수를 구하라.

$$mode = \frac{a-1}{a+b-2} = \frac{1}{3} = \text{모수 } \mu \text{ 추정치} \quad \cdots ①$$

$$Var[x] = \frac{ab}{(a+b)^2(a+b+1)} = 0.2^2 = \frac{1}{25} = \text{추정오차} \quad \cdots ②$$

①, ②를 연립하여 풀면, 치환을 통해서 $b = 2a-1$
$25a^2 - 68a + 28 = (a-2)(27a-14)$
$a > 1$인 해는 $a=2, b=3$

감마분포

감마분포(Gamma distribution)도 베타분포처럼 모수의 베이지안 추정에 사용된다. 다만 베타분포가 0부터 1까지의 사잇값을 베이지안 방법으로 추정하는데 사용되는 것과 달리, 감마분포는 0부터 무한대의 값을 가지는 양수 값을 추정하는 데 사용된다.

~~감마분포의 확률 밀도 함수는 모수 a, b의 값에 따라 다음과 같은 형상을 가진다.~~

감마분포의 확률밀도함수는 a와 b라는 두 모수를 가지며, 수학적으로 다음과 같이 정의된다.

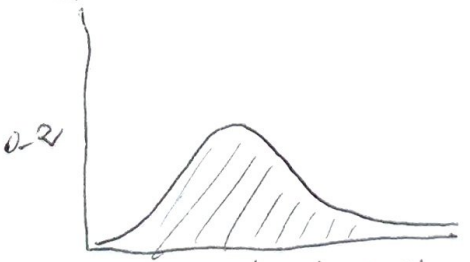
$$Gam(x; a, b) = \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-bx}$$

감마분포의 확률 밀도 함수는 모수 a, b의 값에 따라 다음과 같은 형상을 갖는다.

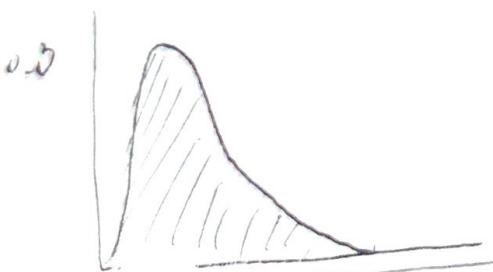
(감마분포의 형상)

(A) a=9, b=1, 최빈값=7

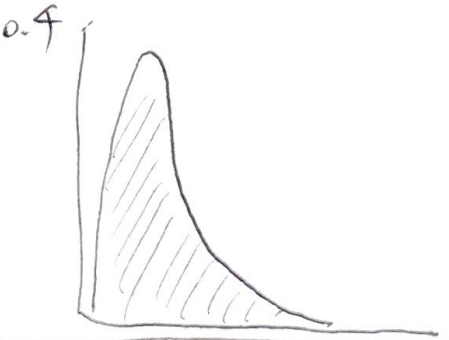
(B) a=6, b=1, 최빈값=5



(C) a=3, b=1, 최빈값=2



(D) a=2, b=1, 최빈값=1



$$E[X] = \frac{a}{b}, \quad mode = \frac{a-1}{b}, \quad Var[X] = \frac{a}{b^2}$$

디리클레분포

디리클레분포(Dirichlet distribution)는 베타분포의 확장판이라고 할 수 있다.
베타분포가 0과 1 사이의 값을 취하는 단일(univariate) 확률변수의
베이지안 모형에 사용되고, 디리클레분포는 0과 1 사이의 값을 갖는
다변수(multivariate) 확률변수의 베이지안 모형에 사용된다.

예를 들어 $K=3$인 디리클레분포를 따르는 확률변수는 다음과 같은 값들을
표본으로 가질 수 있다.

$(0.2, 0.3, 0.5)$
$(0.5, 0.5, 0)$
$(1, 0, 0)$

디리클레분포의 확률밀도함수는 다음과 같다.

베르누이/카테고리분포의 모수 — "모수의 모수" 모수를 결정하는 모수 = hyper parameter

$$Dir(x;\alpha) = Dir(x_1, x_2, \cdots, x_K; \alpha_1, \alpha_2, \cdots, \alpha_K)$$

$$= \frac{1}{B(\alpha_1, \alpha_2 \cdots, \alpha_K)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

이 식에서 $x = (x_1, x_2, \cdots, x_K)$는
디리클레분포의 표본값 벡터이고, $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K)$는 모수벡터다.
$B(\alpha_1, \alpha_2, \cdots, \alpha_K)$는 베타함수라는 특수함수로 다음과 같이 정의한다.

$$B(\alpha_1, \alpha_2, \cdots, \alpha_K) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$$

디리클레분포의 확률값 x는 다음 제한조건을 따른다.

$$0 \le x_i \le 1, \quad \sum_{i=1}^{K} x_i = 1$$

→ 카테고리분포의 모수가 되어야 하므로,

베타분포와 디리클레분포의 관계

베타분포는 $K=2$인 디리클레분포라고 볼 수있다.

$x_1 = x$, $x_2 = 1-x$, $\alpha_1 = a$, $\alpha_2 = b$ 로 하면

$$\begin{aligned} Beta(x; a, b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \\ &= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} \\ &= \frac{1}{B(\alpha_1,\alpha_2)} \prod_{i=1}^{2} x_i^{\alpha_i-1} \end{aligned}$$

디리클레분포의 모멘트

기대값

$$E[x_k] = \frac{\alpha_k}{\sum \alpha}$$

최빈값

$$mode = \frac{\alpha_k - 1}{\sum \alpha - K}$$

모수인 $(\alpha_1, \alpha_2, \dots, \alpha_K)$는 $(x_1, x_2, \dots, x_K)$ 중 어느수가 더 크게 나올 가능성이 높은지를 결정하는 형상인자 (shape factor), 모든 α_i 값이 동일하면 모든 x_i의 분포가 같아진다.

분산

$$Var[x_k] = \frac{\alpha_k(\sum \alpha - \alpha_k)}{(\sum \alpha)^2(\sum \alpha + 1)}$$

$(\alpha_1, \alpha_2, \dots, \alpha_K)$의 절대값이 클수록 분산이 작아진다.
즉, 디리클레 분포의 표본값 x가 어떤 특정한 값 주변이 나올 가능성이 높아진다.

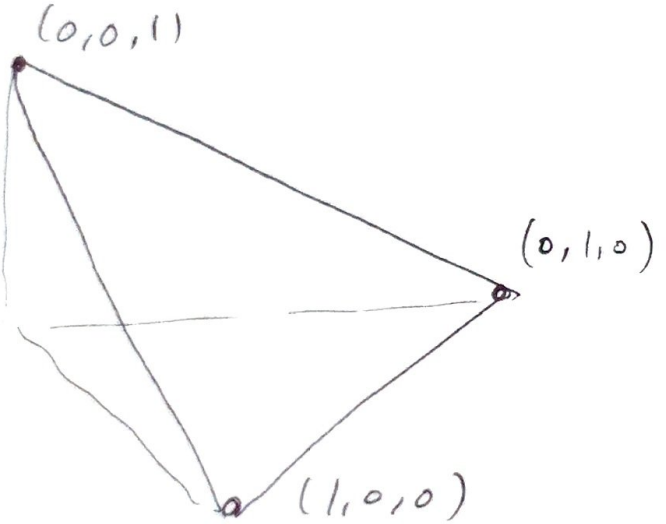
디리클레분포의 응용

다음과 같은 문제를 풀어보자.

{ x, y, z가 양의 실수일때 항상 $x+y+z=1$이 되게 하려면 어떻게 해야될까요? 모든 경우가 균등하게 나와야 합니다. }

⇒ $k=3$이고 $\alpha_1=\alpha_2=\alpha_3$인 디리클레분포의 특수한 경우

⇒ 3차원공간 상에서 $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ 세점을 연결하는 정삼각형 면 위의 점을 생성하는 문제



① $[0,x]$에서 x를 뽑고, $[0,1-x]$에서 y를 뽑고, 나머지를 z

② 난수로 x, y, z를 뽑은 후 $x+y+z$로 각각을 나눠주기

→ ①, ②는 실질적으로 같은 현상발생 (균등한 분포가 나오지 않음)

→ $\alpha_1=\alpha_2=\alpha_3=1$인 디리클레분포를 만들면 됨!

```
sp.stats.dirichlet((1,1,1)).rvs(1000)
```

Scatter plot