확률 분포 2

윤 정 훈

이산형 확률분포

- ✓ 베르누이 분포
- ✓ 이항 분포
- ✓ 포아송 분포
- ✓ 기하 분포
- ✓ 초기하 분포
- ✓ 음이항 분포

베르누이 분포(Bernoulli)

○ 두 가지의 가능한 결과만을 갖는 시행(trial) 또는 실험(experiment)에서 정의되는 분포이다. 결과가 성공이면 1의 값을 가지고, 실패이면 0의 값을 가지는 확률변수를 베르누이 확률변수라고 한다.

✓ 전화가 왔을 때, 전화를 한 사람이 여자인지 남자인지 측정

✓ 주사위를 한 번 던졌을 때, 숫자 2가 나오는지를 측정

○ 확률질량함수

- $f(1) = p$
- $f(0) = 1 - p = q$

○ 기댓값과 분산

- $E(X) = p$
- $Var(X) = p(1 - p)$

이항 분포 (Binomial)

- 확률변수 $X_1, \dots, X_n$이 성공 확률이 p이며 서로 독립인 베르누이 시행으로부터 얻은 것일 때, 그들의 합으로 나타내어지는 $X = \sum_{i=1}^{n} X_i$의 분포이다. $X = \sum_{i=1}^{n} X_i$는 n회의 독립인 베르누이 시행에서 구한 '성공'의 횟수를 의미한다. 확률변수 X가 이항 분포를 따르는 것을 $X \sim B(n, p)$로 표현한다.
  - ✓ 한 축구 선수가 패널티킥을 차면 5번 중 4번을 성공한다고 한다. 이 선수가 10번의 패널티킥을 차서 7번 성공할 확률은?
  - ✓ A회사는 스마트폰의 한 부품을 만드는 회사로, 이 A사의 불량률은 5%로 알려져 있다. 이 회사의 제품 20개를 조사했을 때, 불량이 2개 이하로 나올 확률은?
- 확률질량함수
  - $f(x) = \binom{n}{x} p^x q^{n-x}$ (단, $x = 0, 1, \dots, n$)
- 기댓값과 분산
  - $E(X) = np$
  - $Var(X) = np(1-p)$

이항 분포 예제

- 어떤 주머니에 r개의 빨간 공과 w개의 하얀 공(단, $r+w=N$)이 들어 있다. 이제 n개의 공을 무작위 복원추출하였을 때

  ✓ 각 시행에서 빨간 공을 추출할 확률 :

  $$p=\frac{r}{N}$$

  ✓ X : 추출된 n개의 공 가운데 빨간 공의 개수

  ✓ X의 확률밀도함수 :

  $$f(x)=\binom{N}{n}\left(\frac{r}{N}\right)^{x}\left(1-\frac{r}{N}\right)^{n-x}\text{ , (단, }x=0,1,2,\dots,n)$$

다항 분포 (multinomial)

○ k개의 상호배반인 사건 A_i에 대하여, $\bigcup_{i=1}^{k} A_i = S$이며, $P(A_i) = p_i \ (i = 1, \dots, k \ , \ \sum_{i=1}^{k} p_i = 1)$이다. 실험을 n회 독립 반복시행했을 때 사건 A_i가 일어날 횟수를 $X_i \ (0 \le x_i \le n \ , \ \sum_{i=1}^{k} x_i = n)$라고 하면, 확률벡터 $\mathbb{X} = (X_1, X_2, \dots, X_k)$는 다항분포를 가진다고 한다.

○ $X_1, \dots, X_k$의 결합 확률밀도함수

- $f(x_1, \dots, x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$

포아송 분포

시간이 지남에 따라 일어나는 어떤 특정한 사건의 발생횟수를 고려하자.

- 특정 시간동안 고속도로 구간에서 일어나는 대형사고의 횟수
- 특정 시간동안 전화교환대에 걸려오는 전화의 수
- 특정 시간동안 주유소에 주유를 위한 차량이 도착한 횟수
- How many pennies will I encounter on my walk home?
- How many children will be delivered at the hospital today?
- How many products will I sell after airing a new television commercial?
- How many defects will there be per 100m of rope sold?
- How many mosquito bites bites did you get today after having sprayed with insecticide?

포아송 분포

The Poisson distribution is very similar to the binomial distribution. We are examining **the number of times an event happens**.

What is the difference?

- Whereas the binomial distribution looks at how many times we register **a success over a fixed the total number of trials**,
- the Poisson distribution measures how many times **a discrete event occurs, over a period of continuous space or time**

The random variable X which counts the number of events can take on any nonnegative integer value.

λ : a parameter which represents the **average** or **expected** number of events to happen within experiment (단위시간당 사건의 발생 확률과 의미 동일)

포아송 분포

포아송 분포의 수학적인 정의

- 사건의 발생 횟수에 대해 다음의 세 가지 성질이 성립한다고 가정하자.
  - $\frac{\lim_{h\to 0} P(\text{길이가 } h \text{인 구간 내에서 사건 } A\text{가 } 1\text{회 일어남})}{h} = \lambda$ (즉, 단위시간당 사건의 발생 확률)
  - $\frac{\lim_{h\to 0} P(\text{길이가 } h \text{인 구간 내에서 사건 } A\text{가 } 2\text{회 이상 일어남})}{h} = 0$
  - 서로 겹치지 않는 두 구간 내에서 사건 A의 발생 횟수는 서로 독립이다.
- 주어진 시간 $t > 0$에 대하여 확률변수 $X(t)$를 구간 $[0, t]$내에서 사건 A가 발생하는 횟수라고 하자. X가 아래와 같은 확률밀도함수를 가지고 사건의 발생 횟수에 대한 세 가지 성질을 만족하면, 모수가 λ인 포아송(Poisson) 확률변수라 하고 $X \sim POI(\lambda)$로 표기한다.

포아송 분포

○ X의 확률밀도함수

- $f(x) = \frac{\exp(-\lambda t)(\lambda t)^x}{x!}$ $\left(t = 1,\ f(x) = \frac{\exp(-\lambda)(\lambda)^x}{x!}\right)$

○ 기대값과 분산

- $E(X) = \lambda$
- $Var(x) = \lambda$

포아송 분포 예제 1

○ 어떤 전화교환대에서 매분 평균적으로 2건의 통화가 이루어진다고 한다. 통화횟수가 포아송 확률과정을 따른다는 가정하에, 3분 동안에 5건 이상의 통화가 이루어질 확률은?

- 3분 동안에 걸려온 통화횟수 : 포아송 확률변수 X
- $\lambda = E(X) = 6$이다.
- $P(X \geq 5) = 1 - P(X \leq 4) = 1 - \sum_{x=0}^{4} \frac{6^x e^{-6}}{x!} = 1 - 0.285 = 0.715$

포아송 분포 예제 2

- 어떤 책의 페이지당 오자의 수는 평균 3개라고 한다. 책의 한 페이지를 임의로 펼칠 때 오자의 수가 4개 이상일 확률을 구하여라.

기하 분포

○ "성공"확률이 p인 베르누이 시행을 독립적으로 반복할 때, 첫 번째 "성공"이 일어날 때까지 걸리는 시행횟수(X)를 나타내는 분포이다.

○ X의 확률질량함수

- $f(x) = p(1-p)^{x-1}$ $(x = 1,2, \ldots)$

○ X의 기대값과 분산

- $E(X) = \sum_{x=1}^{\infty} xp(1-p)^{x-1} = \frac{1}{p}$
- $Var(X) = \frac{1-p}{p^2}$

○ 기하확률변수는 무기억성을 갖는다.

- $X \sim GEO(p)$이면, 임의의 자연수 j, k에 대하여 $P(X > j+k | X > j) = P(X > k)$

기하 분포 예제 1

○ 매주 발행되는 어떤 복권의 당첨확률이 0.001이고, 복권이 당첨되는 사건들은 독립이라 하자. 어떤 복권구입자가 매주 이 복권을 살 때, 처음 당첨될 때까지 소요되는 구매횟수를 확률변수 X라고 하자.

- X의 확률질량함수
  - ▷ $f(x) = 0.001(1-0.001)^{x-1}$, $(x = 1,2,\dots)$
- X의 기대값
  - ▷ $E(X) = \frac{1}{0.001} = 1000$
- X의 분산
  - ▷ $Var(X) = \frac{1-0.001}{(0.001)^2} = 999000$

기하 분포 예제 2

- 한 개의 주사위를 계속해서 던지는 실험에서 정확히 10번째에 6의 숫자가 처음으로 나타날 확률을 구하여라.

초기하 분포

○ 어떤 주머니에 r개의 빨간 공과 w개의 하얀 공(단, $r+w=N$)이 들어 있고, 그 중에서 n개의 공을 무작위 비복원추출(sampling without replacement)하였을 때, 선택된 빨간 공의 개수를 확률변수 X라고 하자. 확률변수 X의 확률밀도함수가 아래와 같을 때, X의 분포는 초기하 분포이다.

○ X의 확률밀도함수

- $f(x)=\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$, $x=\max(0,n-N+r),\ldots,\min(r,n)$

○ X의 기댓값과 분산

- $E(X)=\frac{nr}{N}$
- $Var(X)=\frac{nr}{N}\left[\frac{(N-r)(N-n)}{N(N-1)}\right]$

초기하 분포 예제

○ 어떤 상자에 100개의 전자제품이 포장되어 있다. 한 구입자가 이 상자에 10개의 부품을 비복원 랜덤추출하여 불량품의 상태를 조사하려고 한다. 실제로 100개들이 상자속에 30개의 불량품이 있을 때, 2개 이하의 불량품이 나올 확률을 구해 보자.

- 확률변수 X : 불량품의 개수
- $X \sim HYP(10,100,30)$
- $P(X \leq 2) = \sum_{x \leq 2} \frac{\binom{30}{x}\binom{70}{10-x}}{\binom{100}{10}} = 0.372857$

음이항 분포

○ "성공"확률이 p인 베르누이 시행을 독립적으로 반복할 때, r개의 "성공"을 얻을 때까지 필요한 시행 횟수를 확률변수 X라고 하자. X의 확률밀도함수가 아래와 같을 때, 변수 X의 분포는 음이항 분포이다.(negative binomial)

○ X의 확률질량함수

- $f(x;r,p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$, $(x = r, r+1, \dots)$

○ X의 기댓값과 분산

- $E(X) = \frac{r}{p}$
- $Var(X) = \frac{rq}{p^2}$

음이항 분포 예제

○ 두 개의 팀 A와 B가 겨루는 7회의 동일한 게임으로 구성된 경기에서 4회를 먼저 이기는 팀이 우승을 하게 된다. 어떤 팀이든지 먼저 4회를 이기면 경기는 더 이상 계속되지 않고 종료된다. A팀이 각 게임에서 이길 확률을 0.7이라고 할 때, 경기가 5회에서 종료될 확률을 구해 보자.

✓ A팀이 5회째 경기에서 4번째 승리를 거둘 사건 : A_5

✓ B팀이 5회째 경기에서 4번째 승리를 거둘 사건 : B_5

- $P(A_5) = \binom{4}{3}(0.7)^4(0.3) = 0.28812$
- $P(B_5) = \binom{4}{3}(0.3)^4(0.7) = 0.02268$
- 따라서 경기가 5회에서 종료될 확률은 0.28812+0.02268=0.3108이다.

연속형 확률분포

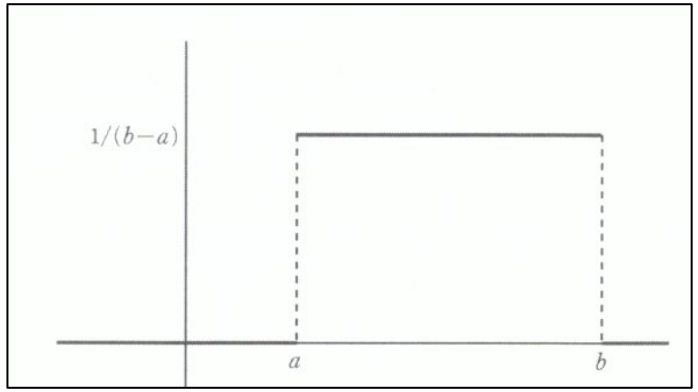
- ✓ 균일 분포
- ✓ 지수 분포
- ✓ 정규 분포
- ✓ 감마 분포
- ✓ 베타 분포

균일 분포

○ 확률변수 X가 실구간 (a,b)상에 균일(uniform)하게 분포되어 있을 때, 아래의 확률밀도함수를 가진다. X가 균일 분포를 따르면 $X \sim U(a,b)$로 표기한다.

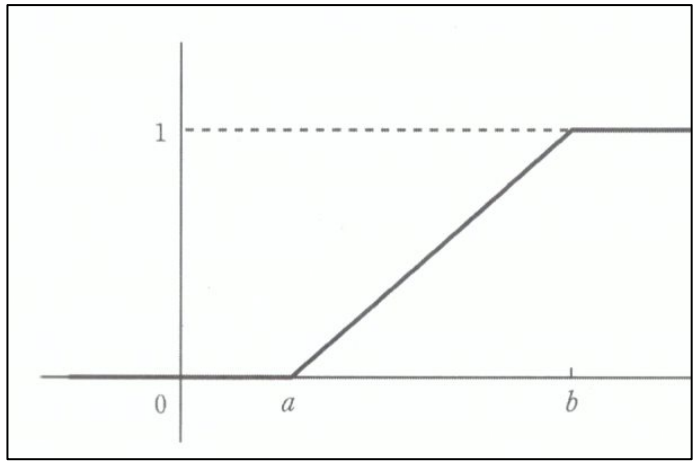
○ X의 확률밀도함수

- $f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{그 외의 경우} \end{cases}$



○ X의 분포누적함수

- $F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x < b \\ 1, & x \ge b \end{cases}$



○ X의 기댓값와 분산

- $E(X) = \frac{b+a}{2}$
- $Var(X) = \frac{(b-a)^2}{12}$

지수 분포

○ 지수분포는 포아송 확률변수를 고려했을 때, 특정한 사건 A가 일어나고 다음에 또 다시 같은 사건이 일어날 때까지 걸리는 시간 X(음이 아닌)을 나타내는 분포이다. 단위구간 내에서 평균발생횟수가 θ인 포아송 과정의 사건 사이 소요시간의 평균값은 $\lambda = 1/\theta$가 된다.

○ X의 확률밀도함수

- $f(x) = \frac{1}{\lambda}\exp(-\frac{x}{\lambda})$, $(x > 0)$

○ X의 누적분포함수

- $F(x) = 1 - \exp(-\frac{x}{\lambda})$

○ X의 기댓값과 분산

- $E(X) = \lambda$
- $Var(X) = \lambda^2$

지수 분포 예제

○ 어떤 화학물질에서 α-입자가 10초당 평균 5개씩 포아송 과정을 따라 발생한다고 하자. 이때 첫 번째 입자가 발생될 때까지 걸리는 시간 X가 5초 이상일 확률을 구해보자.

✓ 초당 발생될 α-입자의 수의 기댓값은 $\lambda = 2$ 즉 첫번째 발생의 기대 소요시간은 $\theta = \frac{1}{2}$

✓ X의 확률밀도함수 :

$$f_X(x) = \left(\frac{1}{2}\right) e^{-\frac{x}{2}}, \ (x \geq 0)$$

✓ X가 5초 이상일 확률 :

$$P(X \geq 5) = \int_5^{\infty} \left(\frac{1}{2}\right) e^{-\frac{x}{2}} dx = e^{-\frac{5}{2}} = 0.08208$$

지수 분포 무기억성

○ 지수분포의 무기억성

- $X \sim EXP(\lambda)$이면, 양의 실수 a와 t에 대해서, $P(X > a + t | X > a) = P(X > t)$
- 예를 들어, 확률변수 X가 어떤 기계부품의 수명이라고 하면, $P(X > a + t | X > a)$는 시점 a에서 기계부품의 고장이 없을 때, 최소한 시간 t만큼 더 고장이 없을 사건에 대한 확률이다. 무기억 성질은 변수 X가 시점 a에서 기계부품의 고장이 없다는 조건을 "기억"하지 않고, 앞으로 시간 t만큼 더 고장이 없을 것만 고려한다는 것을 뜻한다. 즉, a시간만큼 일한 기계부품이 앞으로 t시간만큼 더 작동하는 확률이나 새 기계부품이 앞으로 t시간만큼 더 작동하는 확률이나 같다는 의미이다.

○ 신뢰성이론, 생존모형, 보험계리모형에 사용된다.

정규 분포

○ 통계적 방법에서 가장 많이 이용되는 확률분포이며, 확률변수 X의 확률밀도함수가 다음과 같이 주어지면 X가 정규 분포(normal distribution)를 따른다고 한다.

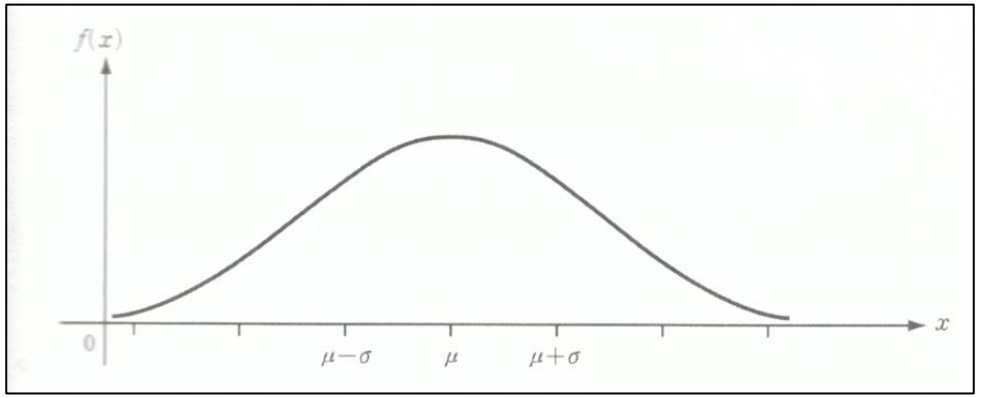
○ 확률밀도함수 :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{(x-\mu)^2}{2\sigma^2}]$$

○ X의 기대값과 분산

- $E(X) = \mu$
- $Var(X) = \sigma^2$

○ X가 정규분포를 따르면 $X \sim N(\mu, \sigma^2)$로 표기하고, $Y = aX + b$에 대하여, $Y \sim N(a\mu + b, a^2\sigma^2)$로 나타낼 수 있다.



표준 정규 분포

- 정규분포를 따르는 확률변수 X에 대하여 , $\boldsymbol{Z = \frac{X-\mu}{\sigma}}$라는 표준화 변환을 하면, Z의 확률밀도함수는 다음과 같이 주어진다. 이것을 표준 정규 분포(standard normal distribution) 라고 한다.
- 확률밀도함수 :

  $$f_Z(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})$$

- Z가 표준 정규 분포를 따르면 $Z \sim N(0,1)$이라고 한다.
- Z의 누적분포함수 :

  $\Phi(x) = P(Z \leq x) = \int_{-\infty}^{x} \phi(t)dt$ 라고 할 때, X의 누적분포함수는 $F(x) = \Phi(\frac{x-\mu}{\sigma})$

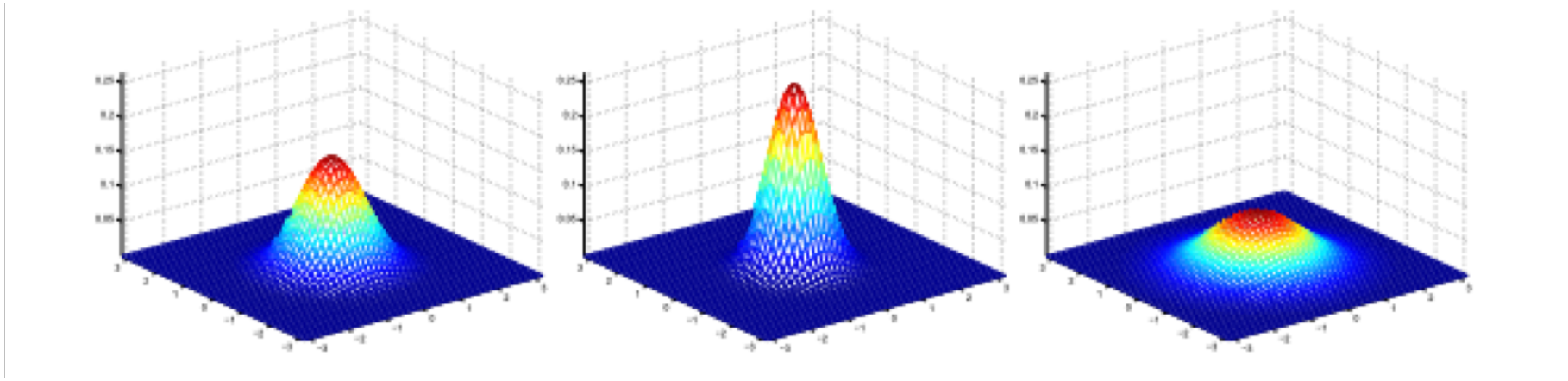
이변량 정규 분포

- 두 변수 X,Y 의 결합 정규 분포 (bivariate normal distribution)이다.
- X의 결합 확률 밀도 함수 :

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

- X가 이변량 정규 분포를 따를 경우, $X \sim BVN(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$라고 표현한다.
- X, Y의 기댓값, 분산
  - $E(X) = \mu_X$, $E(Y) = \mu_Y$, $Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$
- 공분산
  - $Cov(X,Y) = \rho\sigma_X\sigma_Y$

이변량 정규 분포 예제



(Andrew Ng, Generative learning algorithm lecture note)

Tails of normal distribution

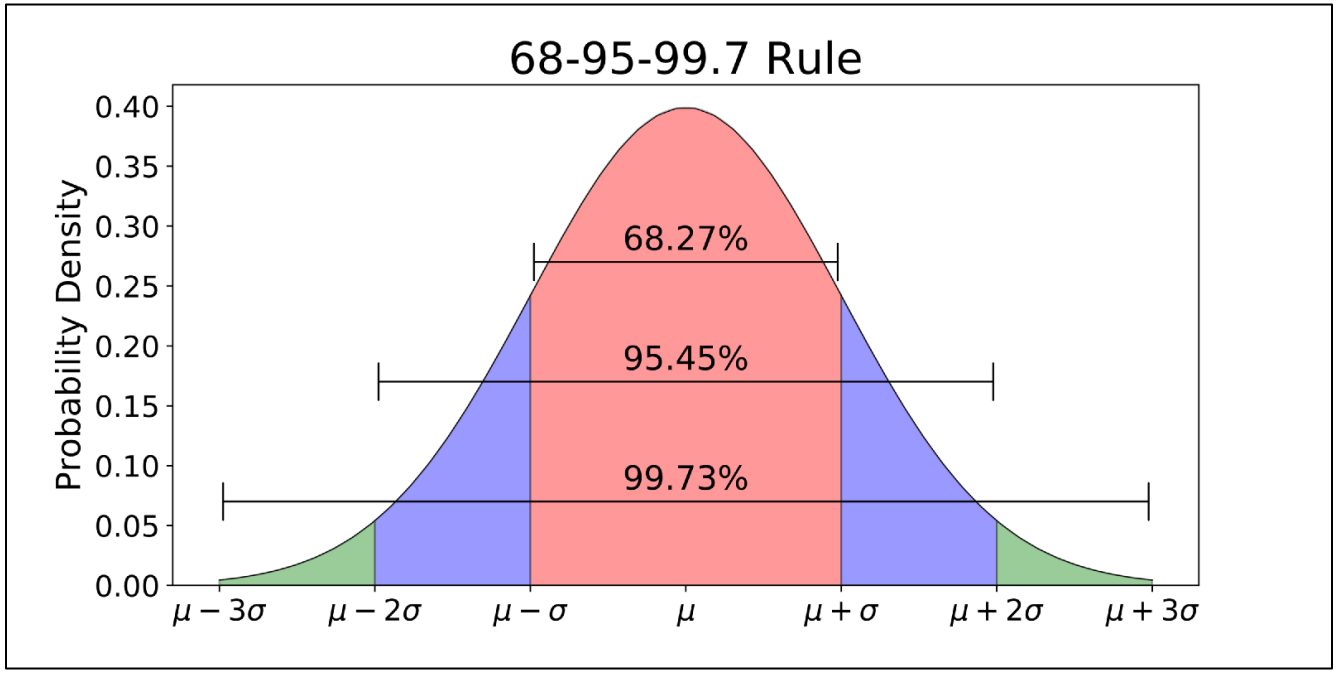


그림 출처 : 구글

범위	확률	
	범위 안	범위 밖
$\mu \pm \sigma$	68.3%	31.7%
$\mu \pm 2\sigma$	95.4%	4.6%
$\mu \pm 3\sigma$	99.7%	0.27%

베타 분포

- 베타 분포는 데이터의 분포를 묘사하는 것 뿐 아니라 다른 확률분포함수의 모수를 베이지안 추정(Bayesian estimation) 한 결과를 표현하기 위해 사용된다. 모수가 가질 수 있는 모든 값에 대한 가능성을 확률분포로 나타낸 것이다.
- 베르누이 분포는 2가지의 결과값 0 or 1을 가지고 있다. 1이 발생할 확률이 p이면 0이 발생할 확률은 $1-p$이다. p는 모수이며, 어떠한 값이라도 가질 수 있다. p가 가질 수 있는 0 부터 1사이의 값에 대한 가능성은 바로 **베타 분포**로 표현 할 수 있다.

베타 분포

○ 베타 함수 :

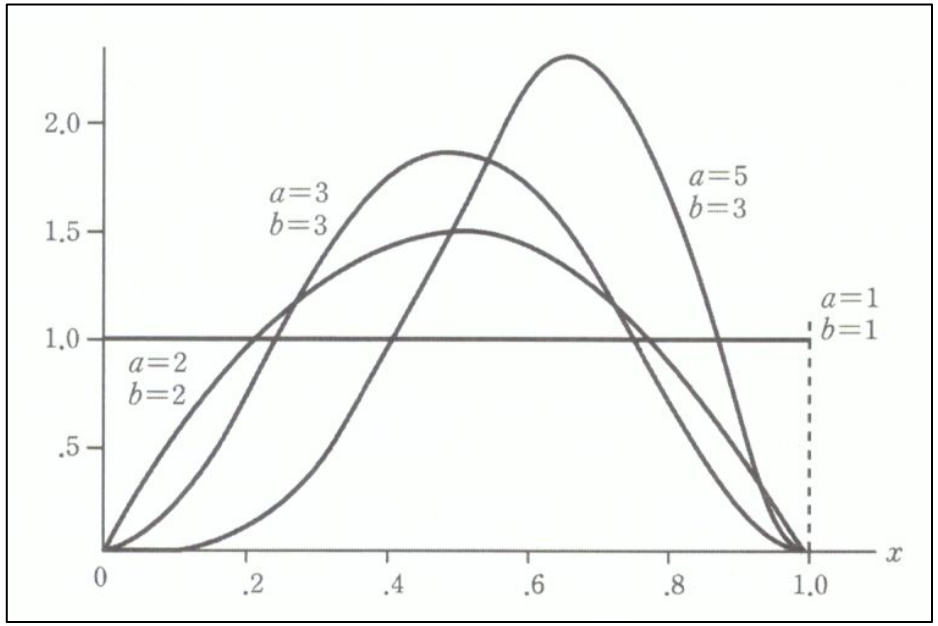
$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dt$$

○ 확률변수 X가 다음의 확률밀도함수를 가지면, 베타분포라고 한다. 단 x의 값은 $0 < x < 1$ 이다.

$$f(x;\theta,k) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$$

○ X의 기댓값과 분산

- $E(X) = \frac{a}{a+b}$
- $Var(X) = \frac{(a+1)a}{(a+b+1)(a+b)} - \left(\frac{a}{a+b}\right)^2 \left(\frac{ab}{(a+b+1)(a+b)^2}\right)$
- Mode(최빈값) = $\frac{a-1}{a+b-2}$



감마 분포

- 지수분포 등의 여러 가지 분포를 포함하는 분포족(family of distributions)의 하나이다. 감마 분포를 알기 위해서는 감마 함수를 알아야 한다.
- 감마 함수
  - $k > 0$에 대하여, $\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt$
  - 감마함수의 특징
    - ✓ $\Gamma(1) = \int_0^\infty e^{-t} dt = 1$
    - ✓ $\Gamma(k) = (k-1)\Gamma(k-1)$, $k > 1$에 대하여
    - ✓ $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

감마 분포

○ 감마분포는 모수의 베이지안 추정에 사용된다. $0 \sim \infty$ 의 값을 가지는 양수 값을 추정할 때 사용된다.

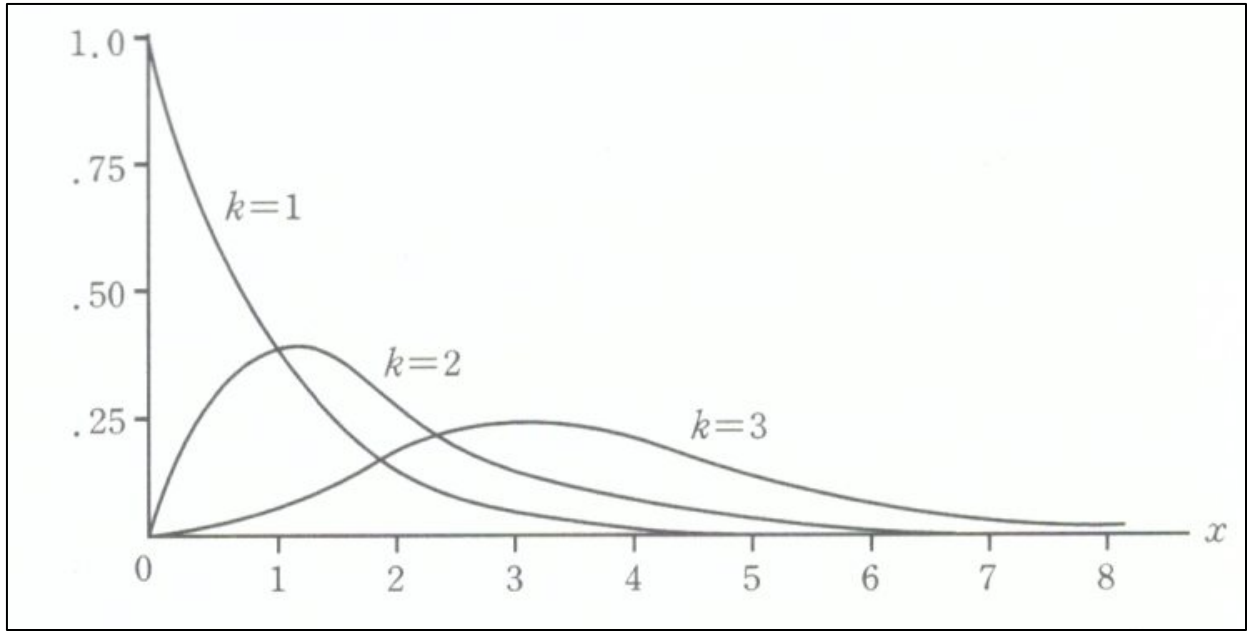
○ $\theta > 0$, $k > 0$, $x > 0$ 에 대하여, 확률변수 X가 아래의 확률밀도함수를 가지면, 감마분포라고 한다.

○ X의 확률밀도함수 :

$$f(x;\theta,k) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} \exp(-\frac{x}{\theta})$$

○ X의 기댓값과 분산

- $E(X) = k\theta$
- $Var(X) = k\theta^2$
- Mode(최빈값) $= \frac{a-1}{b}$



켤레 사전 분포

베르누이 분포의 매개변수 p에 대해서 생각해보자.

이항 분포에서는 p의 최대 가능도 추정값(MLE)이 데이터에 있는 $x = 1$ 의 관측값의 비율로써 계산된다.

그렇다면 데이터의 수가 적을때는? 이 관측값만 사용할 수 있을까? 과적합이 일어나기 쉽다. 따라서 Bayesian으로 접근해보자.

Bayesian 으로 접근하기 위해서는 매개변수에 대한 사전 분포 $p(p)$를 도입하고자 한다.

해석하기 쉽고, 분석 측면에서 유용한 형태의 사전 분포를 도입하려고 한다.

베르누이의 가능도 함수는 $p^x(1-p)^{1-x}$ 의 형태를 가지는 인자들의 곱 형태이다.

p 와 $(1-p)$의 거듭제곱에 비례하는 형태를 사전 분포로 선택한다면, 사전 확률과 가능도 함수의 곱에 비례하는 사후 분포 역시 사전 분포와 같은 형태를 가지게 될 것이다.

이러한 성질이 켤레성(conjugacy)이다. 베르누이의 사전 분포는? 베타 분포이다.

켤레 사전 분포

- 다항 분포의 켤레 분포는 디리클레 분포이다. 디리클레 분포는 왜 중요한가? LDA!
- **가우시안 분포의 평균에 대한 켤레 분포는 가우시안 분포이다. (정밀도가 알려져 있다는 가정하에)
- **가우시안 분포의 정밀도(분산의 역수)에 대한 켤레 분포는 위샤트 분포이다. (평균이 알려져 있다는 가정하에)