2016/03/04

[2016-동계세미나-확률 기초]

Chapter 6 연속형 균일분포 (2)

- 6.4~6.10 -

이 부 형(boohyung@pel.smuc.ac.kr)

상명대학교 프로토콜공학연구실

Contents

- 이항분포의 정규근사
- 감마분포와 지수분포
- 카이제곱분포
- 베타분포
- 로그정규분포
- 와이블 분포

이항분포의 정규근사

- 이항분포의 정규근사: 이항분포에서 n이 충분히 크다면 정규분포의 면적을 이용하여 이항분포의 근사값을 계산할 수 있는 원리
  - 이항분포에서 n이 크고, p가 아주 작거나 크지 않은 경우
  - 이항분포에서 n이 작고, p가 ½에 가까운 값일 경우

- X가 $\mu = \mathrm{np}$이고 $\sigma^2 = npq$인 이항확률변수이면, $n \to \infty$일 때

$$Z = \frac{X - np}{\sqrt{npq}}$$

는 표준정규분포, 즉 $n(z; 0, 1)$을 따름

이항분포의 정규근사

- 연속성 수정: 이항분포의 정규근사에서, 이항분포는 이산형 분포이기 때문에 연속형 분포로 더 적합하게 근사시켜 확률을 정확하게 구하기 위해 수정하는 값
  - $X \pm 0.5$
- X를 모수 n과 p를 갖는 이항확률변수라고 하자. 그러면 X는 평균이 $\mu = np$이고 분산이 $\sigma^2 = npq$인 정규분포를 근사적으로 따르게 됨

$$P(X \le x) = \sum_{k=0}^{x} b(k; n, p) \approx P(Z \le \frac{x + 0.5 - np}{\sqrt{npq}})$$

  - 이항분포의 정규근사는 np와 nq가 5 이상일 때 더 적합
    - ref. 교재에서 표 6.1 정규근사와 실제누적이항확률(표)

이항분포의 정규근사

- 예제 6.15) 빈혈환자가 회복될 확률은 0.4라고 한다. 100명의 빈혈환자 중에서 회복되는 환자의 수가 30보다 적을 확률은?

  풀이) 이항변수 X는 회복될 환자의 수, $n = 100$

  $$\mu = np = 100(0.4) = 40$$

  $$\sigma = \sqrt{npq} = \sqrt{100(0.4)(0.6)} \approx 4.899$$

  연속성 수정에 의해 $x = 30 - 0.5 = 29.5$

  $$z = \frac{29.5-40}{4.899} = -2.14$$

  $$P(X < 30) \approx P(Z < -2.14) = 0.0162$$

이항분포의 정규근사

- 예제 6.16) 4지선다형 문제 200개가 있다고 하자. 그 시험에 대해 전무한 학생이 200문제 중 80문제의 답을 추측만으로 골랐을 때, 정답이 25개에서 30개일 확률은?

  풀이) X가 그 학생이 맞춘 정답의 수, $p = 0.4$

$$P(25 \le X \le 30) = \sum_{x=25}^{30} b(x; 80, \frac{1}{4})$$

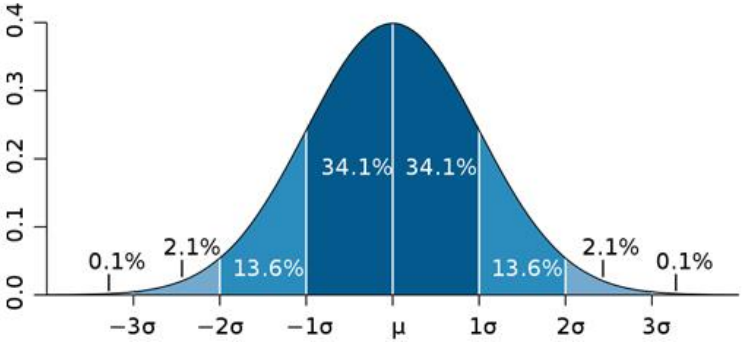
$$\mu = np = 80\left(\frac{1}{4}\right) = 20$$

이항분포의 정규근사

- 예제 6.16) 4지선다형 문제 200개가 있다고 하자. 그 시험에 대해 전무한 학생이 200문제 중 80문제의 답을 추측만으로 골랐을 때, 정답이 25개에서 30개일 확률은?

풀이) $\sigma = \sqrt{npq} = \sqrt{80\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)} \approx 3.873$

- 표준편차의 의미
  - 표본의 68.2%: 16.127~23.873
  - 표본의 95.4%: 12.254~27.746



이항분포의 정규근사

- 예제 6.16) 4지선다형 문제 200개가 있다고 하자. 그 시험에 대해 전무한 학생이 200문제 중 80문제의 답을 추측만으로 골랐을 때, 정답이 25개에서 30개일 확률은?

  풀이) 이항분포를 정규곡선에 근사시키면

  $x_1 = 24.5, x_2 = 30.5$

  $$z_1 = \frac{24.5 - 40}{3.873} = 1.16, \qquad z_2 = \frac{30.5 - 40}{3.873} = 2.71$$

  $$P(25 \le X \le 30) = \sum_{x=25}^{30} b(x; 80, 0.25)$$

  $$\approx P(1.16 < Z < 2.71) = 0.9966 - 0.8770 = 0.1196$$

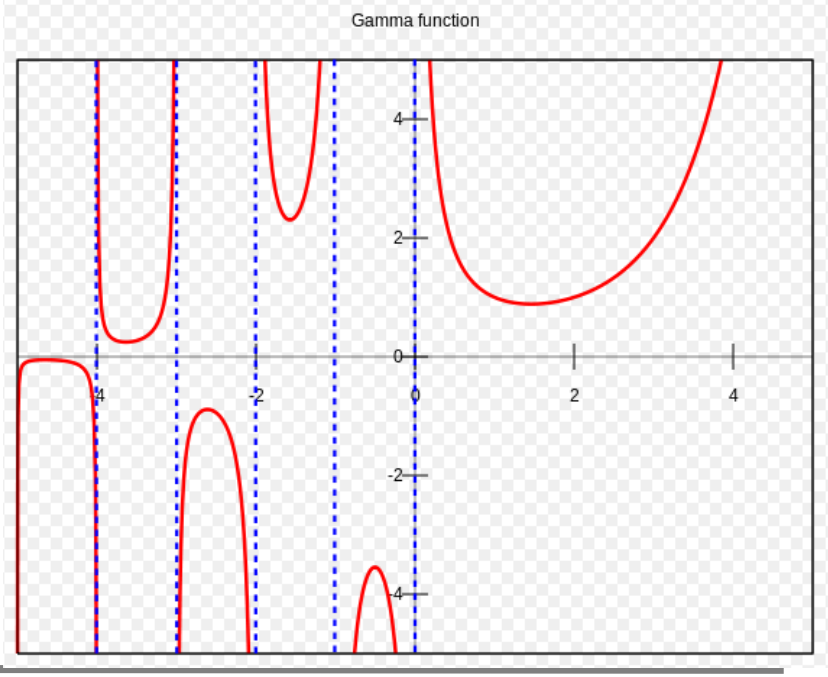
감마분포와 지수분포

- 감마함수(factorial 함수의 정의역 확장: 복소수까지)

$\alpha > 0$인 α에 대해서

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

- $\Gamma(n) = (n-1)(n-2)\ldots(1)\gamma(1) = (n-1)!$
- $\Gamma(1) = 1$
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$



감마분포와 지수분포

- 감마분포: 특정한 횟수만큼의 포아송 사건이 발생할 때까지의 시간을 나타내는 분포
  - 특정한 횟수가 모수 α

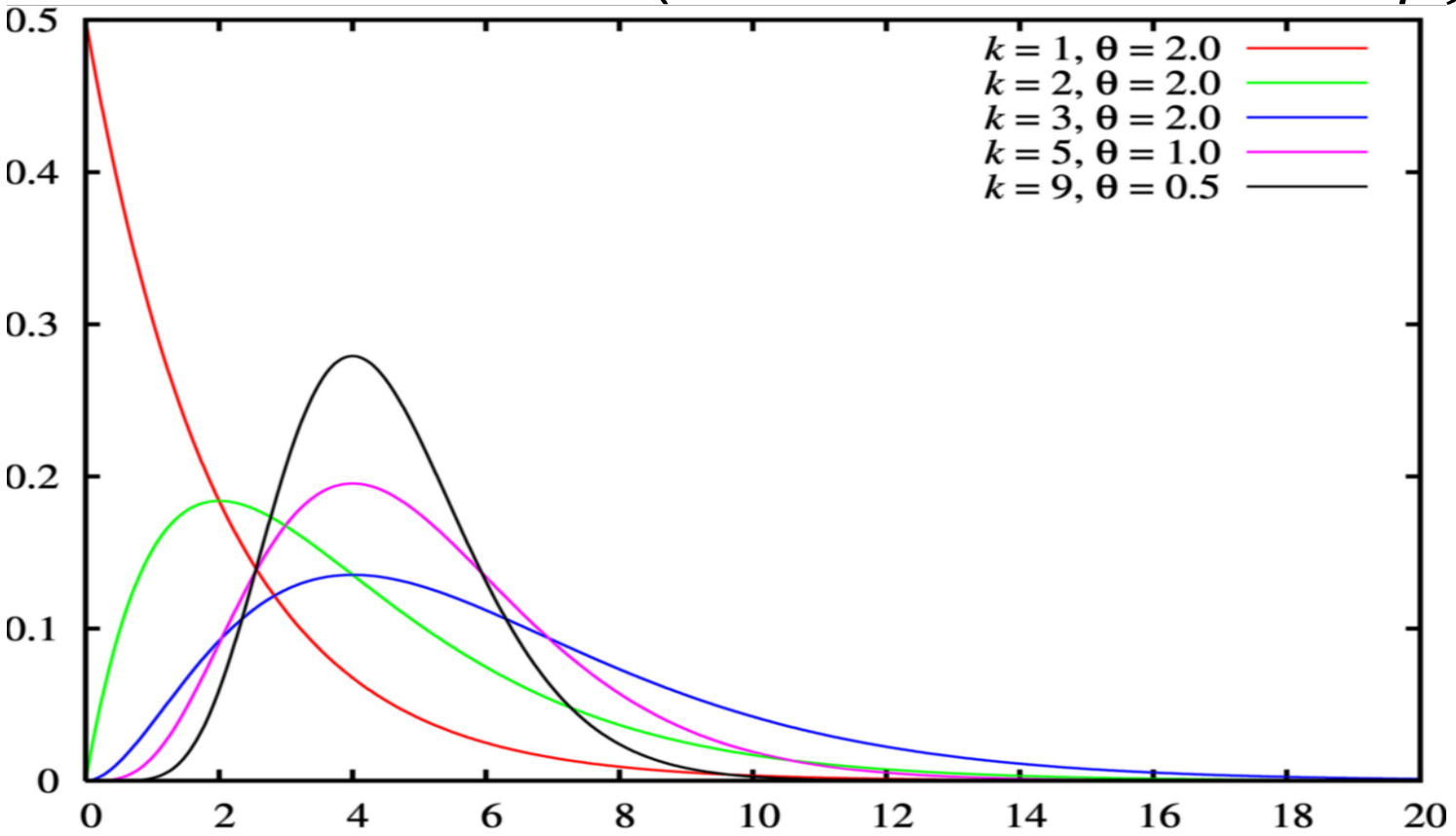
- 연속확률변수 X의 확률밀도함수가

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\beta}}, & x>0\\ 0, & \text{다른 곳에서}(\text{단},\alpha>0,\beta>0)\end{cases}$$

과 같이 주어질 때, X는 모수 α,β를 가지는 감마분포를 따름

감마분포와 지수분포

- 감마분포의 그래프(그래프에서 $k = \alpha, \theta = \beta$)



- 감마분포의 평균과 분산

$$\mu = \alpha\beta, \qquad \sigma^2 = \alpha\beta^2$$

감마분포와 지수분포

- 지수분포: $\alpha = 1$인 감마분포

  연속형 확률변수 X의 밀도함수가

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & \text{다른 곳에서(단, } \beta > 0) \end{cases}$$

  과 같이 주어질 때, X는 모수 β를 가지는 지수분포를 따름

- 포아송 분포(이산) → 지수분포(연속)
  - 어느 사건이 포아송분포를 따른다면, 이러한 사건들이 발생하는 사이에 경과되는 시간은 지수분포를 따름

감마분포와 지수분포

- 포아송 분포(이산) → 지수분포(연속)
  - 어느 사건이 포아송 분포를 따른다면, 이러한 사건들이 발생하는 사이에 경과되는 시간은 지수분포를 따름
  - 예제(네트워크)
    - 라우터에 패킷 도착: 포아송 분포
    - 라우터에서 패킷이 머무르는 동안의 처리 시간: 지수분포

감마분포와 지수분포

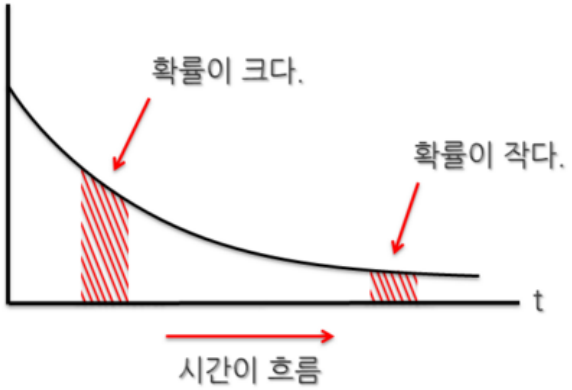
- 지수분포의 예
  - 전자제품이나 기계가 시간이 지날수록 고장 나지 않을 확률
  - 시간이 지날수록 휴대폰 배터리가 남아 있을 확률
  - 사람이 나이를 먹을수록 살아있을 확률

- 지수분포의 평균과 분산

$$\mu = \beta, \qquad \sigma^2 = \beta^2$$

  - λ를 이용한 평균과 분산, 표준편차 계산

평균= $\frac{1}{\lambda}$, 분산= $\frac{1}{\lambda^2}$, 표준편차= $\sqrt{\frac{1}{\lambda^2}}$



감마분포와 지수분포

- 예제 6.18) 전화교환기에 도착되는 호출신호는 분당 평균이 5회인 포아송 과정을 따른다고 한다. 1분 내에 2번의 호출신호가 도착될 확률은?

  풀이) 2번의 호출신호가 도착되기까지의 소요 시간을 X라고 하면, 이는 2번의 포아송 사건이 발생되기까지 소요된 시간

  $$\alpha = 2, \qquad \beta = \frac{1}{5}$$

  $$P(X \le 1) = \int_0^1 \frac{1}{\beta^2} x e^{-\frac{x}{\beta}} dx = 25 \int_0^1 x e^{-5x} dx$$

  $$= 1 - e^{-5}(1 + 5) = 0.96$$

카이제곱분포

- 카이제곱분포
  - 분산이 퍼져있는 모습을 분포로 만든 것
  - 감마분포의 두 번째 특수한 경우
    - $\alpha = \frac{\nu}{2}, \beta = 2$ (ν: 자유도)
  - 카이제곱분포의 정의: 표준정규 확률변수 X의 확률분포가

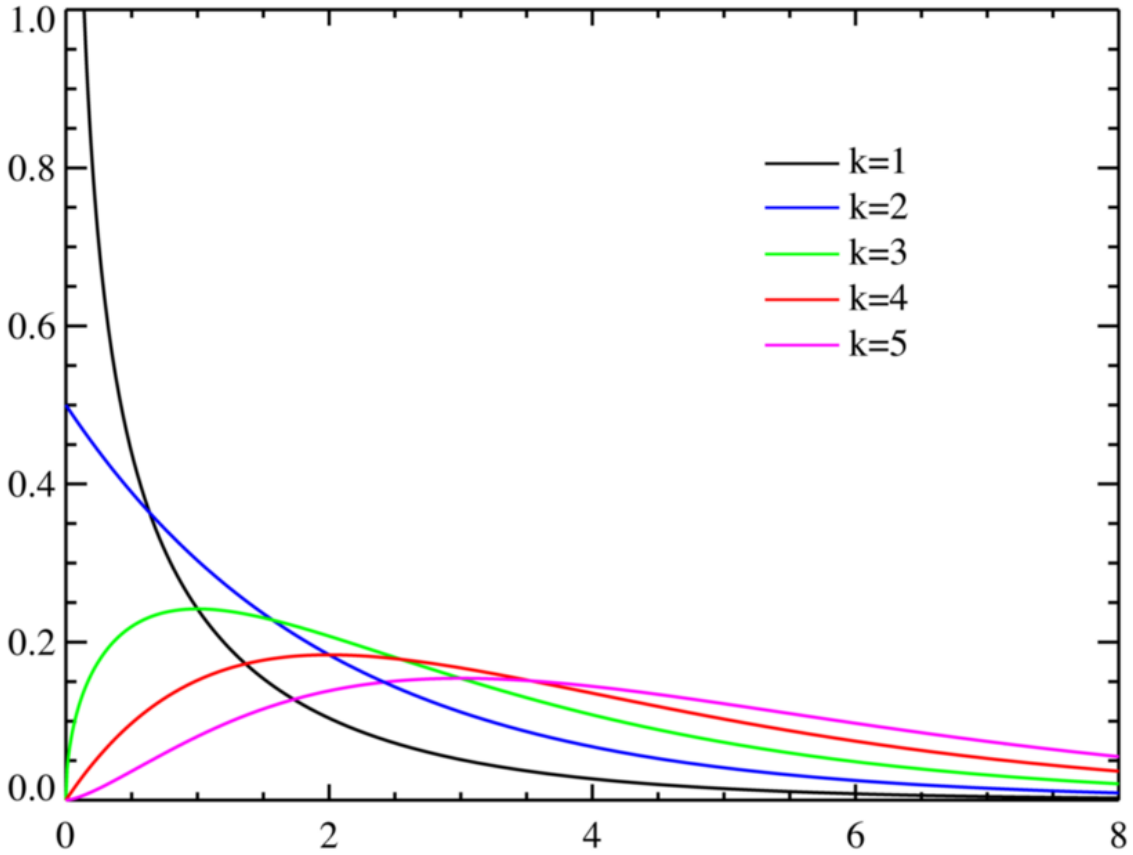
$$f(x;\nu) = \begin{cases} \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{\frac{-x}{2}}, x > 0 \\ 0, \text{다른 곳에서(단, } \nu\text{는 양의 정수)} \end{cases}$$

  와 같이 주어질 때, X는 자유도 ν인 카이제곱분포를 따른다고 함
  - 카이제곱분포의 평균과 분산

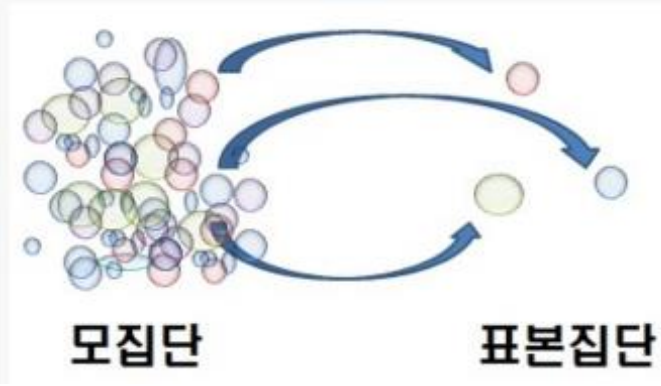
$$\mu = \nu, \qquad \sigma^2 = 2\nu$$

카이제곱분포

- 카이제곱분포
  - 카이제곱분포의 그래프
    - $k = v$(자유도)



카이제곱분포

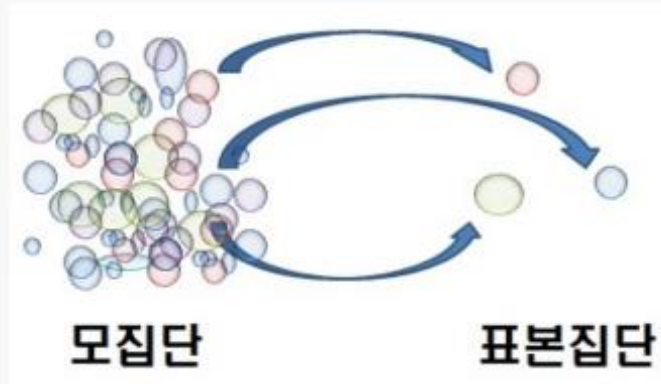


- 카이제곱분포
  - 자유도의 의미
    - 주어진 조건에서 자유롭게 뽑을 수 있는 수
      - 표본의 수 - 1
    - 예제> 대한민국 전체 고등학교 3학년 학생의 수학 점수의 평균이 50점이다. 이 때, 몇몇 대표할만한 학생들의 수학 성적을 바탕으로 정보를 추출하려고 한다. 표본의 수는 100명으로 정하고 평균은 50점이다.

      → 99명은 임의로 뽑을 수 있지만 나머지 1명은 모집단의 평균과 표본집단의 평균은 같아야 하므로 임의로 뽑을 수 없음 (∵ 표본집단은 모집단을 대표한다)

      자유도 ν = 100 - 1 = 99

카이제곱분포



- 카이제곱분포
  - 자유도의 의미
    - 주어진 조건에서 자유롭게 뽑을 수 있는 수
      - 표본의 수 - 1
    - 예제> 대한민국 전체 고등학교 3학년 학생의 수학 점수의 평균이 50점이다. 이 때, 몇몇 대표할만한 학생들의 수학 성적을 바탕으로 정보를 추출하려고 한다. 표본의 수는 100명으로 정하고 평균은 50점이다.

      → 100명의 표본을 추출할 때, 만약 99명의 점수 합이 4,928점이라면, 나머지 1명을 뽑을 때는 무조건 점수가 72점인 학생을 뽑아야 함
      (∵ 100명의 총 점수 합 = 5,000)

카이제곱분포

- 카이제곱분포
  - 이용: 카이제곱 검정
    - 두 변수의 연관성에 대한 검정
      - 각 범주에 대한 기대값을 구함
      - 범주별 카이제곱 값 구하기
        - 카이제곱 값 = (관측값-기대값)^2/기대값
      - 전체 카이제곱 값 구하기
      - 자유도 구하기
      - 유의수준에 해당하는 카이스퀘어 값과 비교하여 결론
        - 유의수준(p-value): 귀무가설을 기각할 수 있는 최소한의 확률
        - 이 예제에서 귀무가설은 “흡연과 주량은 연관성이 없다.”
      - 예제. 흡연량과 음주량 사이에는 연관성이 있는가?

카이제곱분포

- 카이제곱분포
  - 이용: 카이제곱 검정
    - 두 변수의 연관성에 대한 검정
      - 예제. 흡연량과 음주량 사이에는 연관성이 있는가?

	1갑 이상	1갑 이하	안피움	계
반 병 이상	23	21	63	107
반 병 이하	31	48	159	238
못마심	13	23	119	155
계	67	92	341	500

      - 범주별 기대값 구하기
        - 1갑 이상 + 반 병 이상은 23명
        - 67(1갑 이상의 계)*107(반 병 이상의 계)/500(총계) = 14.338

카이제곱분포

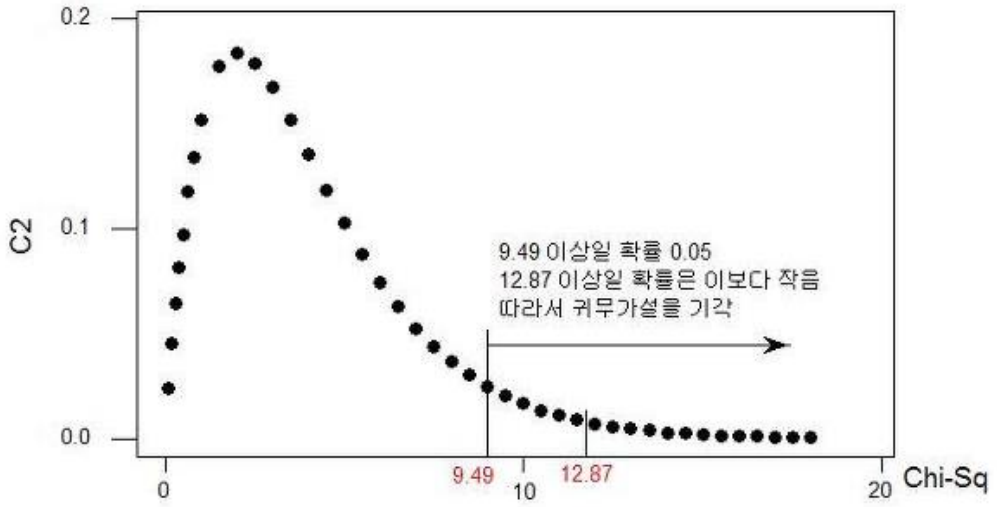
- 카이제곱분포
  - 이용: 카이제곱 검정
    - 두 변수의 연관성에 대한 검정
      - 예제. 흡연량과 음주량 사이에는 연관성이 있는가?
      - 범주별 카이제곱 값 구하기

$$\chi^2 = \frac{(\text{관측값} - \text{기대값})^2}{\text{기대값}}$$

      - 전체 카이제곱 값 구하기(예제에서 전체 카이제곱 값은 12.87)
      - 흡연량과 음주량의 자유도는 각각 2; 카이제곱 검정을 위한 자유도는 2 * 2 = 4

카이제곱분포

- 카이제곱분포
  - 이용: 카이제곱 검정
    - 두 변수의 연관성에 대한 검정
      - 예제. 흡연량과 음주량 사이에는 연관성이 있는가?



    - 카이제곱분포표를 확인하여 기준 카이제곱 값을 결정
    - 서로 비교하여 계산한 카이제곱 값이 크다면, 귀무가설을 기각
    - 결과: 음주량과 흡연량은 연관성이 있음

베타분포

- 베타분포
  - 균일분포의 일반화
  - 베타함수($\alpha, \beta > 0$이고 $\Gamma(\alpha)$는 감마함수)

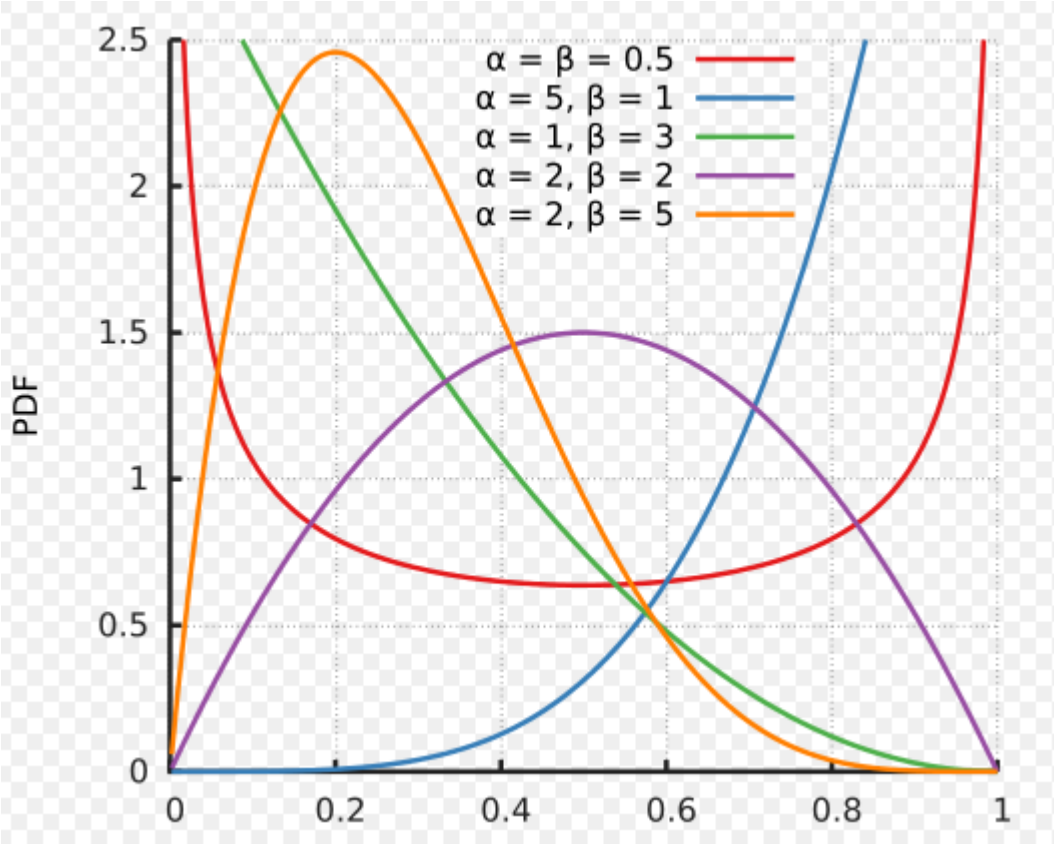
$$\mathrm{B}(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

  - 정의: 연속형 확률변수 X의 밀도함수가 다음과 같을 때 확률변수 X는 모수가 $\alpha > 0$이고 $\beta > 0$인 베타분포를 따름

$$f(x) = \begin{cases} \dfrac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1 \\ 0, \text{다른 곳에서} \end{cases}$$

베타분포

- 베타분포
  - 베타분포의 그래프



베타분포

- 베타분포
  - 베타분포의 평균과 분산

$$\mu = \frac{\alpha}{\alpha + \beta}, \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

  - 베타분포의 활용
    - 제조과정에서의 불량품의 비율(불량율)
    - 화학약품의 불순률
    - 기계의 가동률

베타분포

- 베타분포
  - 예제> 공급받은 제품중 불량율을 나타내는 확률변수 X는 $B(6,3)$ 을 따른다. 불량인 제품이 80%이상일 확률은?

풀이) $B(6,3) = \frac{\Gamma(6)\Gamma(3)}{\Gamma(9)} = \frac{1}{168}$

$$f(x) = 168x^5(1-x)^2, 0 < x < 1$$

$$P(X \geq 0.8) = \int_{0.8}^{1} 168x^5(1-x)^2 dx$$

$$= 168\left[\frac{x^6}{6} - \frac{2x^7}{7} + \frac{x^8}{8}\right]_{0.8}^{1} = 0.2013$$

로그정규분포

- 로그정규분포

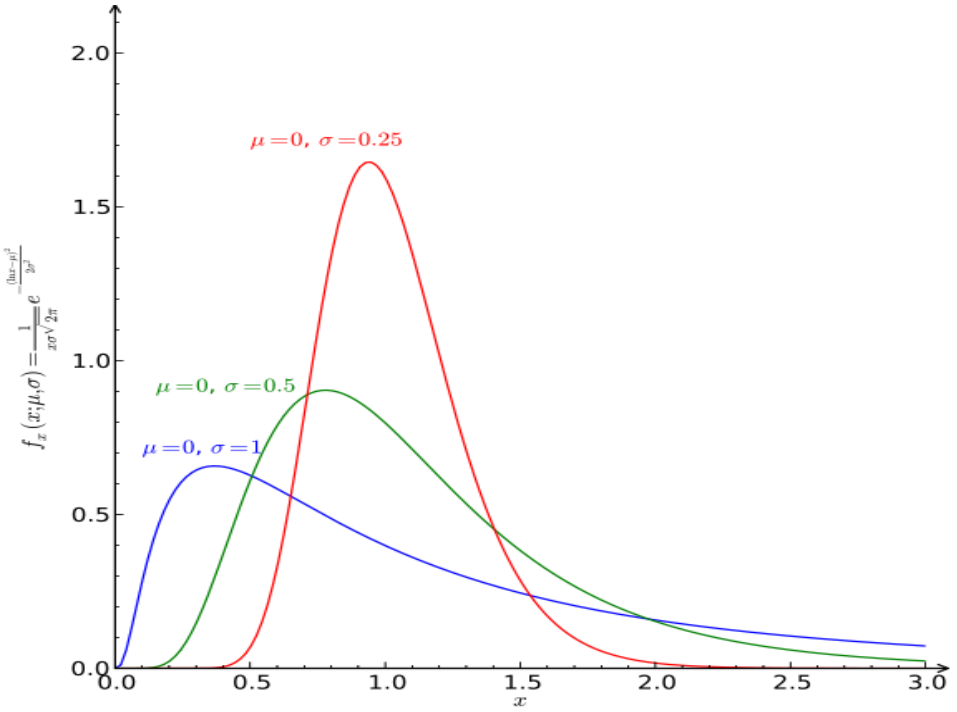
  확률변수 $Y = \ln(x)$가 평균 μ이고 표준편차 σ인 정규분포를 따를 때, 확률변수 X의 분포

$$f(x;\ \mu, \sigma) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

  - 확률변수 x에 자연로그를 취했을 때, 그 결과가 정규분포를 따름
  - 활용: 신뢰도 분석, 보험에서 손해액 추정

로그정규분포

- 로그정규분포
  - 로그정규분포의 그래프



  - 로그정규분포의 평균과 분산

$$\mu = e^{\mu+\frac{\sigma^2}{2}}, \qquad \sigma^2 = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$$

와이블 분포

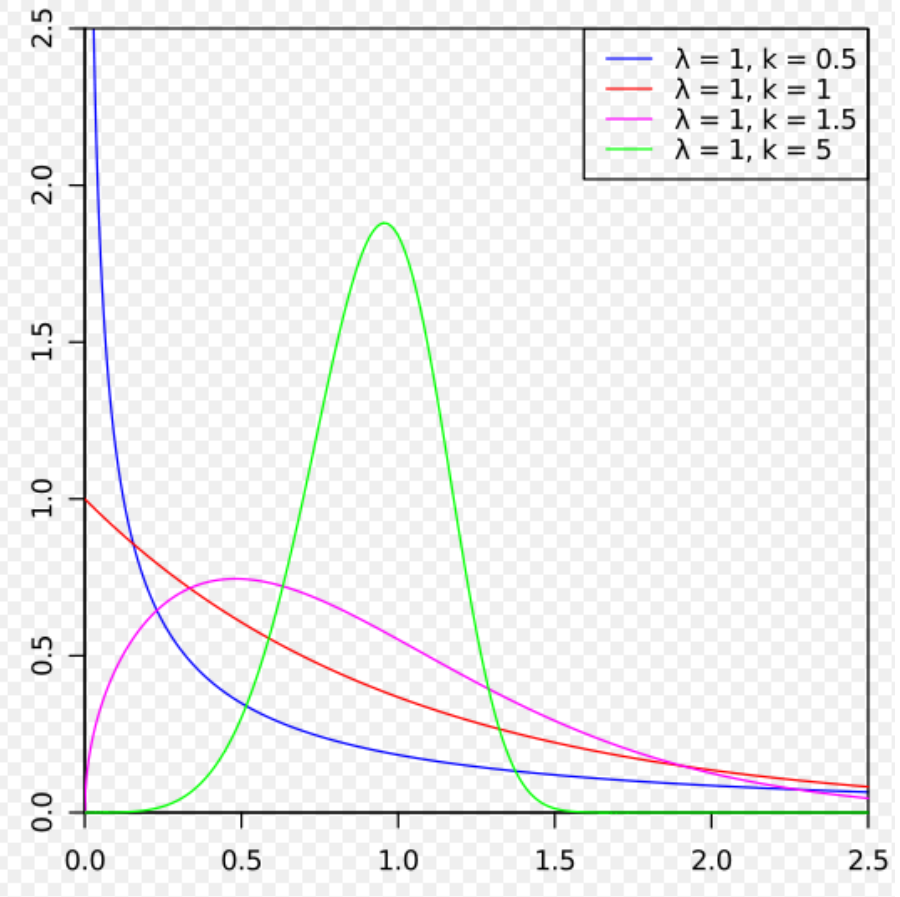
- 와이블 분포
  - 1939년 스웨덴의 물리학자 와이블(Weibull)에 의해 고안됨
  - 활용: 시스템의 작동과 안전에 영향을 주는 부품의 신뢰성 문제
    - e.g., 퓨즈가 타는 문제, 강철이 휘어지는 문제, 열감지센서가 고장나는 문제
    - 동일한 환경에 적용되는 동일한 부품이라 할 지라도 고장시기를 예측하기는 힘듬
  - 연속확률변수 X의 확률분포가

$$f(x;\alpha,\beta) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}, & x > 0 \\ 0, & \text{다른 곳에서}(\text{단}, \alpha > 0, \beta > 0) \end{cases}$$

과 같이 주어질 때, X는 모수 α, β를 가지는 와이블 분포를 따름

와이블 분포

- 와이블 분포
  - 와이블 분포의 그래프(pdf)
    - 앞의 식에서 $\alpha = \frac{1}{\lambda}, \beta = k$
      - $\alpha = 1$로 고정하고 β를 변화시킴
      - $\beta = 1$일 때, 와이블 분포는 지수분포가 됨
      - $\beta > 1$일 경우, 정규곡선과 비슷하나, 대칭을 이루지는 않음



와이블 분포

- 와이블 분포
  - 와이블 분포의 평균과 분산

$$\mu = \alpha^{-\frac{1}{\beta}}\Gamma\left(1+\frac{1}{\beta}\right),\ \sigma^2 = \alpha^{-\frac{2}{\beta}}\{\Gamma\left(1+\frac{2}{\beta}\right) - \left[\Gamma\left(1+\frac{1}{\beta}\right)\right]^2\}$$

  - 와이블 분포의 누적분포함수

$$F(x) = 1 - e^{\alpha x^{\beta}}, x \geq 0, \alpha > 0, \beta > 0$$

와이블 분포

- 와이블 분포
  - 와이블 분포의 고장률: t(시간 t까지 고장 나지 않았다고 할 때)와 $t + \Delta t$사이에 고장날 조건부확률의 변화율

$$Z(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{R(t)} = \frac{F'(t)}{R(t)}$$
$$= \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)} = \alpha\beta t^{\beta - 1}, t > 0$$

  - $R(t)$: 시점 t에서의 주어진 부품의 신뢰도
    $= P(T > t) = \int_t^{\infty} f(t)dt = 1 - F(t)$
    - 신뢰도: 부분이나 생산품이 규정된 조건하에서 최소한 어느 특정시간까지 작동할 확률

와이블 분포

- 와이블 분포
  - 와이블 분포의 고장률
    - 고장률의 의미
      - $\beta = 1$이면 고장률은 α(상수)
      - $\beta > 1$이면 $Z(t)$는 증가함수: 부품이 시간에 따라 마모되는 현상을 표현
      - $\beta < 1$이면 $Z(t)$는 감소함수: 시간이 지남에 따라 오히려 강해지는 것을 의미

와이블 분포

- 와이블 분포
  - 공학분야에서의 사용 예
    - 수명 확률 분포: 신뢰성 공학에서 소프트웨어가 고장나기까지의 수명의 분포
      - 최적의 소프트웨어 방출 시기 지정
        - 소프트웨어 제품을 개발하여 테스팅을 거친 후 사용자에게 인도하는 시기를 결정하는 문제
        - 고장발생 수명분포 계산
      - 소프트웨어의 수명분포모형 모델링
        - NHPP(Non-Homogeneous Poisson Process)
      - 의료장비 오류 발생 주기 예측