확률 및 통계학

- 6장 연속형 확률분포(2) -

명 세인(sein@pel.smuc.ac.kr)

상명대학교 프로토콜공학연구실

목 차

- 이항분포의 정규근사
- 감마분포와 지수분포
- 카이제곱분포
- 베타분포
- 로그정규분포
- 와이블분포

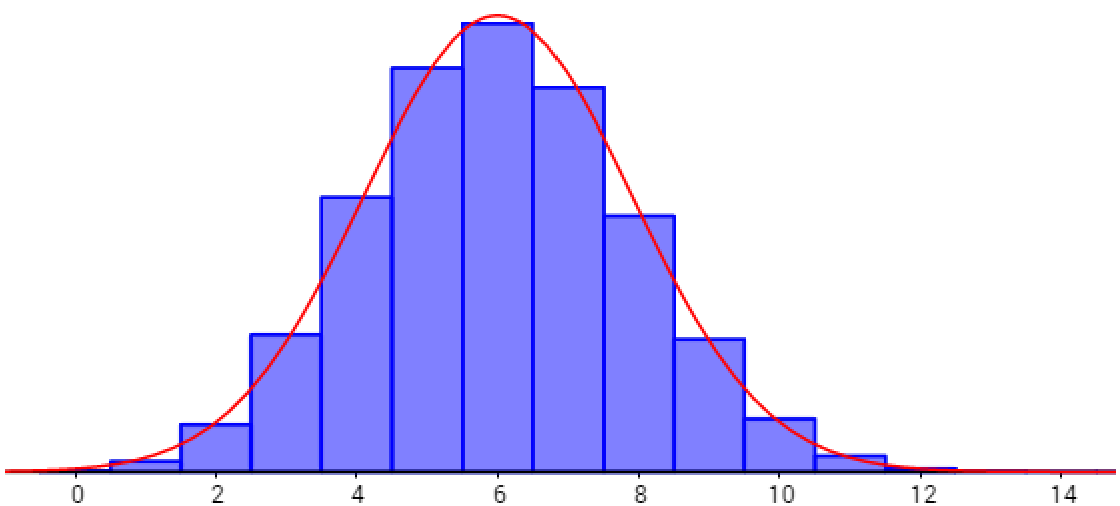
이항분포의 정규근사

- 의미
  - 이상형분포인 이항분포가 시행횟수가 매우 크고 종의 형태처럼 대칭일 때 정규분포에 근사
  - 이항분포를 정규근사하면 계산이 간편해짐

- 정의
  - X가 $\mu = np, \sigma^2 = npq$인 이항확률변수이면, $n \to \infty$일 때
    - $Z = \frac{X-np}{\sqrt{npq}}$ 의 극한분포는 표준정규분포, 즉 $n(z; 0,1)$을 따름

이항분포의 정규근사

- 특징
  - 연속성 수정(Continuity Correction)
    - 이항분포를 정규분포로 근사하여, 특정 값 x의 왼쪽면적을 구하는 경우 이산형과 연속형의 차이를 수정하기 위해 $x + 0.5$를 사용



이항분포의 정규근사

- 특징
  - 이항분포의 정규 근사($n \to \infty$아닐때)
    - X를 모수 n과 p를 갖는 이항확률변수라고 할 때, X는 평균이 $\mu = np$이고 분산이 $\sigma^2 = npq$인 정규분포를 근사적으로 따름
      - $P(X \leq x) = \sum_{k=0}^{x} b(k; n, p) \approx P(Z \leq \frac{x+0.5-np}{\sqrt{npq}})$
      - np와 nq가 5 이상인 경우 근사가 적합

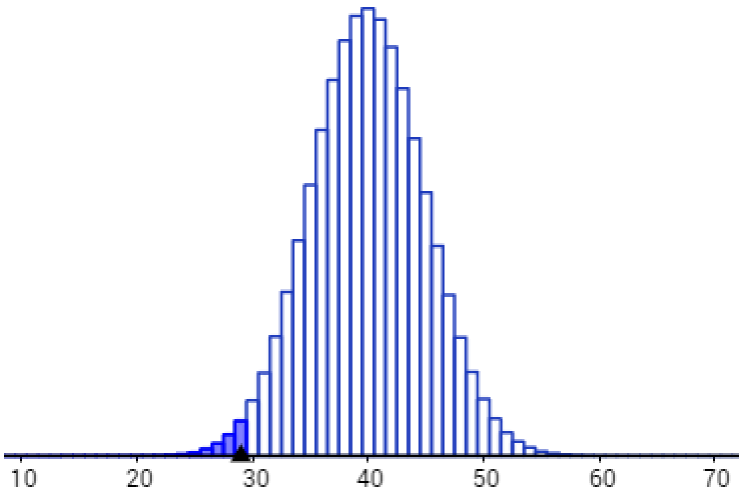
    - n이 아주 큰 값이면 근사 결과가 적합
    - P가 $\frac{1}{2}$에 가까운 값일 때 n이 작아도 근사 결과가 적당

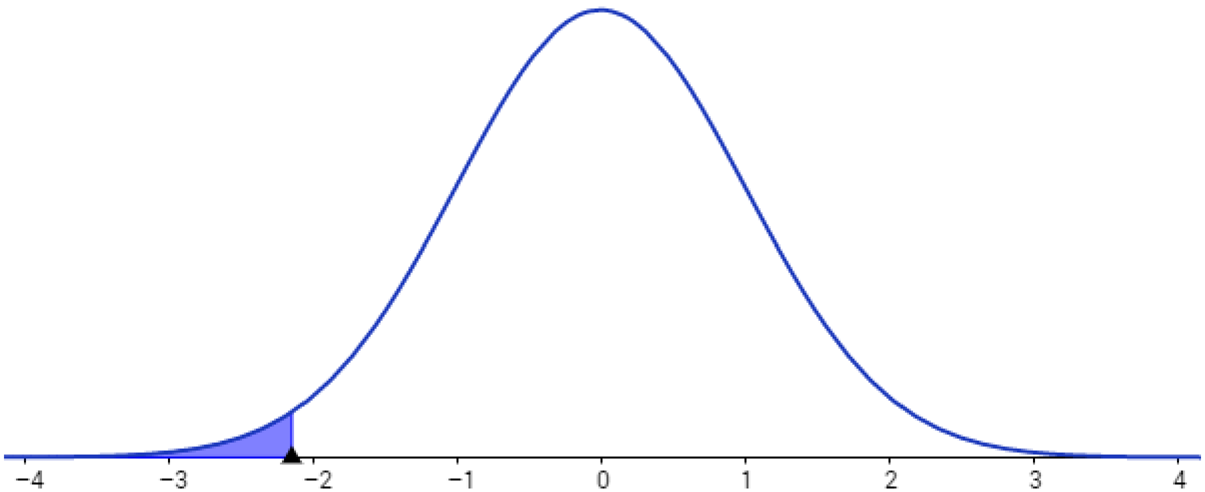
이항분포의 정규근사

- 예제 6.15
  - 빈혈환자가 회복될 확률은 0.4, 100명의 빈혈환자 중 회복되는 환자의 수가 30보다 적을 확률
    - 이항변수 X는 회복될 환자의 수 이고, $n = 100$이므로 정규분포로 근사
      - $\mu = np = (100)(0.4) = 40$
      - $\sigma = \sqrt{npq} = \sqrt{(100)(0.4)(0.6)} = 4.899$
    - 구하고자 하는 확률은 $x = 29.5$의 왼쪽 면적
      - $z = \frac{29.5-40}{4.899} = -2.14$
    - 따라서 100명중 30명보다 적은 사람이 회복될 확률
      - $P(X30) \approx P(Z < -2.14) = 0.0162$
    - 이항누적분포로 계산된 값: $B(29, 100, 0.4) = 0.0148$

이항분포의 정규근사

- 예제 6.15
  - 이항분포 $B(29, 100, 0.4)$와 정규분포$P(Z < -2.14)$



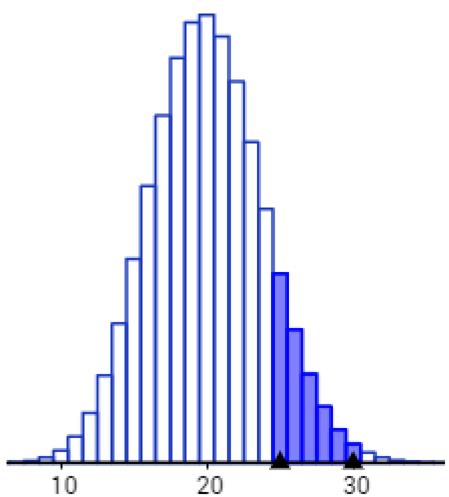


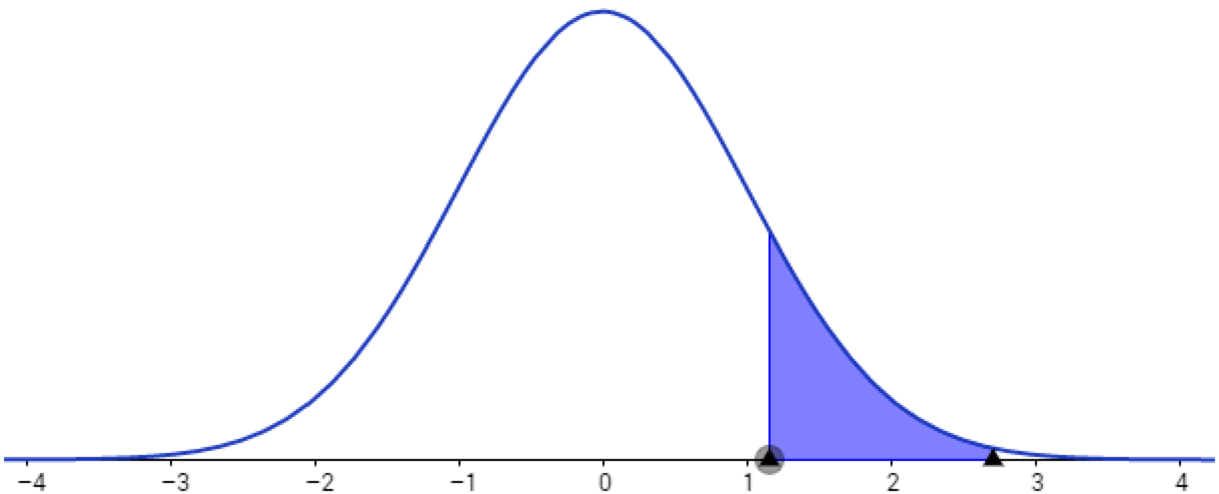
이항분포의 정규근사

- 예제 6.16
  - 4개의 보기 중 하나의 정답이 있는 4지선다형 문제 200개가 있을 때, 200문제 중 80문제의 답을 순전히 추측으로 골랐을 때 25개에서 30개 까지의 확률
    - 80문제에서 답을 할 확률은 $p = 1/4$ 확률변수 X가 맞춘 정답의 수 일때
      - $P(25 \leq X \leq 30) = \sum_{x=25}^{30} b\left(x,\ 80, \frac{1}{4}\right)$
      - $\mu = np = (80)\left(\frac{1}{4}\right) = 20$
      - $\sigma = \sqrt{npq} = \sqrt{(80)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)} = 3.873$ 인 정규분포에 근사
      - $x_1 = \frac{24.5-20}{3.873} = 1.16, x_2 = \frac{30.5-20}{3.873} = 2.71$
      - $P(25 \leq X \leq 30) = \sum_{x=25}^{30} b(x; 80, 0.25) \approx P(1.16 < Z < 2.71)$ $= P(Z < 2.71) - P(Z < 1.16) = 0.9966 - 0.8770 = 0.1196$
      - 이항분포로 구하면 $P(25 \leq X \leq 30) = 0.1193$

이항분포의 정규근사

- 예제 6.16
  - 이항분포$P(25 \le X \le 30)$와 정규분포 $P(1.16 < Z < 2.71)$ 비교





감마분포와 지수분포

- 감마분포
  - 수학의 많은 분야에서 나타나는 감마함수로 유도해낸 확률분포
  - 감마분포의 확률변수: 특정횟수만큼 사건이 발생하기까지의 시간(또는 시간간격)

  - 감마분포의 특수한 경우가 다른 연속형 확률분포로 나타남
    - 카이제곱분포, 지수분포, 와이블분포

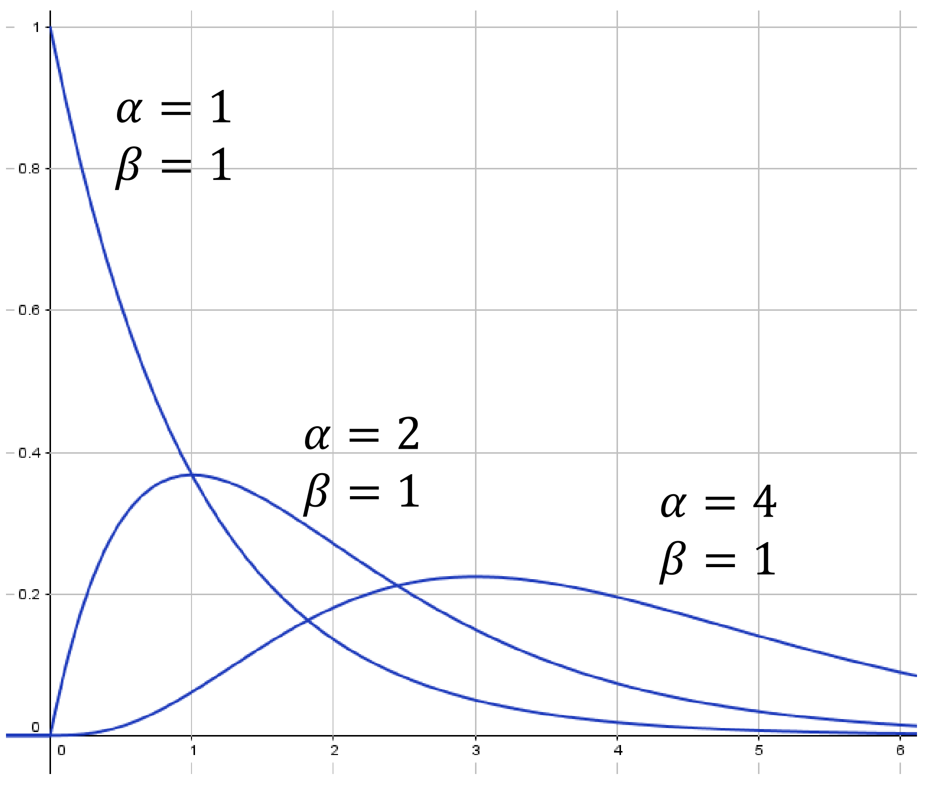
  - 감마함수
    - 자연수 범위의 계승함수(Factorial)를 연속형 함수로 해석
    - 정의
      - $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$

감마분포와 지수분포

- 감마분포
  - 정의
    - 연속확률변수 X의 밀도함수가 다음을 따를 때, X는 모수 α, β를 갖는 **감마분포**를 따름
      - $f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, x > 0 \\ 0, \text{구간 외 (단}, \alpha > 0, \beta > 0) \end{cases}$
  - 특징
    - 감마분포의 평균과 분산
      - $\mu = \alpha\beta$
      - $\sigma^2 = \alpha\beta^2$

감마분포와 지수분포

- 감마분포
  - 특징
    - 모수 α(Shape Parameter)에 따른 감마분포 그래프
      - 모수 β는 Scale Parameter



감마분포와 지수분포

- 지수분포
  - 감마분포의 모수 $\alpha = 1$인 특수한 경우
  - 정의
    - 연속형 확률변수 X의 밀도함수가 다음을 따를 때, 모수 β를 갖는 **지수분포**를 따름
      - $f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}}, x > 0 \\ 0, \text{구간외 (단}, \beta > 0) \end{cases}$
  - 특징
    - 지수분포의 평균과 분산
      - $\mu = \beta$
      - $\sigma^2 = \beta^2$

감마분포와 지수분포

- 지수분포와 포아송 과정과의 관계
  - 포아송분포의 확률변수: **단위시간에 발생횟수**
  - 지수분포의 확률변수: **사건과 사건 사이의 경과 시간**

  - 수식 유도
    - t시간 동안 하나의 사건도 발생하지 않을 확률
      - $p(0;\lambda_t) = \frac{e^{-\lambda_t}(\lambda_t)^0}{0!} = e^{-\lambda_t}$

    - 첫 번째 사건이 발생하기까지 소요된 시간이 x보다 클 확률
      - x시간 내에 포아송사건이 발생하지 않을 확률
      - $P(X > x) = e^{-\lambda x}$
      - X의 누적분포함수는 $P(0 \leq X \leq x) = 1 - e^{-\lambda x}$

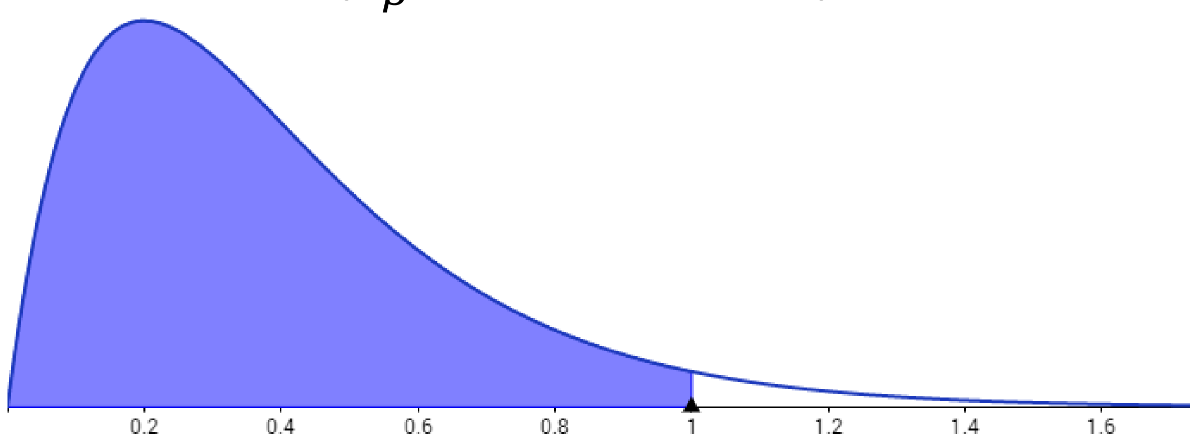
감마분포와 지수분포

- 지수분포와 포아송 과정과의 관계
  - 수식 유도
    - X의 누적분포함수를 미분하면, $\lambda = \frac{1}{\beta}$인 지수확률분포의 확률밀도함수가 나타남
      - $f(x) = \lambda e^{-\lambda x}$
    - 지수분포의 평균을 의미하는 β는 포아송분포의 λ의 역수
  - 포아송 분포의 건망성은 지속적인 시간간격에서 발생하는 사건들이 상호 독립
    - β는 사건간의 평균시간
      - β는 평균고장간격(Mean Time Between Failure)

감마분포와 지수분포

- 예제 6.17 - 지수분포
  - 부품이 고장 나기까지의 시간(단위: 년)을 나타내는 확률변수 T는 고장 나기까지 평균시간이 $\beta = 5$인 지수분포를 따를 때, 이 부품 5개가 각각 다른 시스템에 설치되고, 8년이 지난 후 적어도 2개의 부품이 작동하고 있을 확률
    - 하나의 부품이 8년 이상 작동할 확률
      - $P(T > 8) = \frac{1}{5}\int_8^\infty e^{-\frac{t}{5}} dt = e^{-\frac{8}{5}} \approx 0.2$
    - 확률변수 X는 8년이 지난 후 작동하고 있을 부품의 수, 이항분포를 적용
      - $P(X \geq 2) = \sum_{x=2}^{5} b(x; 5,0.2) = 1 - \sum_{x=0}^{1} b(x; 5,0.2) = 1 - 0.7373$
$= 0.2627$

- 예제 6.18
  - 전화교환기에 도착되는 호출신호는 분당 평균이 5회인 포아송 과정을 따를 때, 1분 내에 2번의 호출신호가 도착될 확률
    - 2번의 호출신호가 도착되기까지 소요된 시간을 X라 하면, 2번의 포아송 사건이 발생되기까지 소요된 시간
      - $\alpha = 2, \beta = \frac{1}{5}$인 감마분포를 따르는 확률변수
      - $P(Z \leq 1) = \int_0^1 \frac{a}{\beta^2} x e^{-\frac{x}{\beta}} dx = 25 \int_0^1 x e^{-\frac{5x}{\beta}} dx = 1 - e^{-5(1+5)} = 0.96$



감마분포와 지수분포

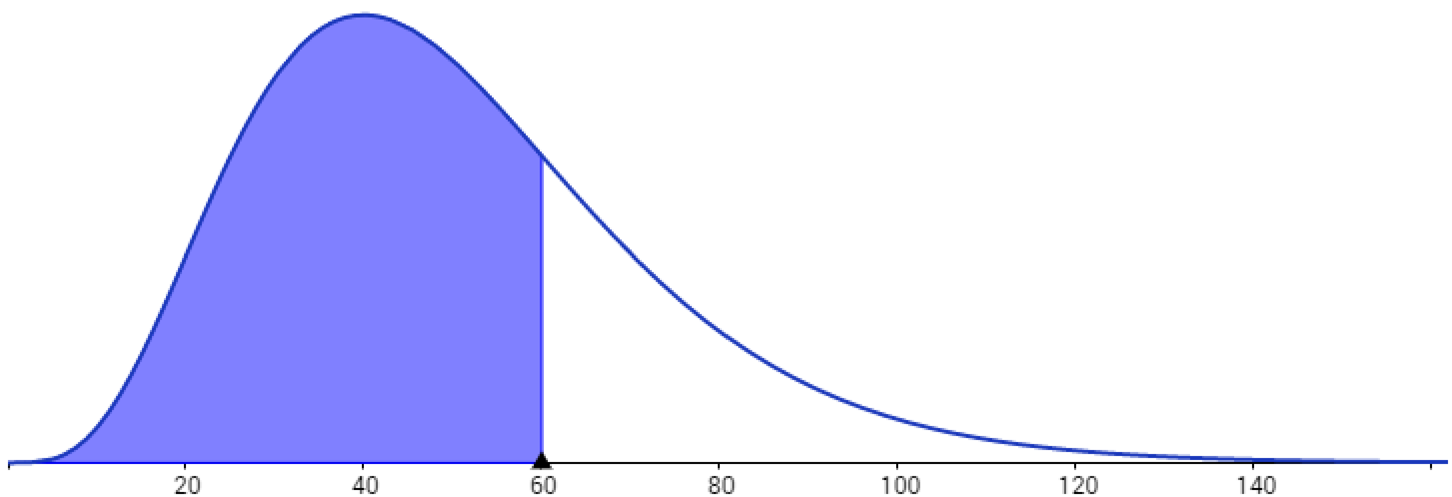
- 지수분포의 건망성
  - 부품의 수명이나 설비의 신뢰성 문제에 사용되는 지수분포는 건망성(Memoryless)특성을 가짐
    - 어떤 전자부품의 수명이 지수분포를 따를 때, 부품의 수명이 t시간 이상 될 확률 $P(X \geq t)$은 다음 특징을 나타냄
      - $P(X \geq t) = P(X \geq t_0 + t \mid X \geq t_0)$
    - 마모가 발생하는 기계부품에는 지수분포 보다는 감마분포나 와이블분포를 적용하는 것이 바람직

감마분포와 지수분포

- 예제 6.19
  - 쥐에 대한, 일정량의 독극물에 대한 생존시간(단위: 주)은 $\alpha = 5$이고 $\beta = 10$인 감마분포를 따를 때, 어떤 쥐에 독극물을 주입하고 60주 이상 생존하지 못할 확률
    - 생존시간을 확률변수 X라고 하면
      - $P(X \leq 60) = \frac{1}{\beta^5}\int_0^{60} \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(5)}dx$
    - 불완전 감마함수(Incomplete Gamma Function) 적분
      - $F(x;\alpha) = \int_0^x \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}dy$
      - 식에서 $y = \frac{x}{\beta}, x = \beta y$로 두면
      - $P(X \leq 60) = \int_0^6 \frac{y^4 e^{-y}}{\Gamma(5)}dy = F(6;5) = 0.7150$
        - $F(x;\alpha) = \int_0^x \frac{1}{\Gamma(\alpha)}y^{\alpha-1}e^{-y}dy$ →불완전 감마함수표 로 나타남(A.22)

감마분포와 지수분포

- 예제 6.19
  - $P(X \leq 60)$=0.7150

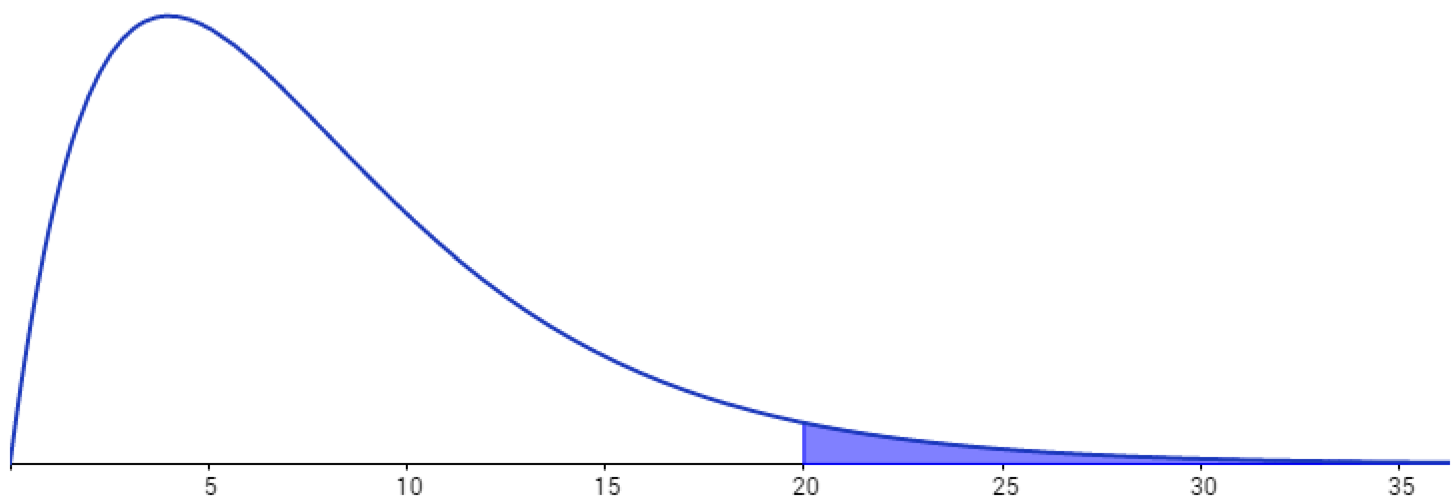


감마분포와 지수분포

- 예제 6.20
  - 제품에 대한 고객불만 제기의 시간간격(단위: 개월)은 $\alpha = 2,\ \beta = 4$ 인 감마분포를 나타내었다. 최근 품질관리 후 첫 불만이 발생할 때 까지 20개월이 소요되었을 때, 품질관리가 효과적이었는가?
    - 불만제기의 시간을 X라고 할 때, $X \geq 20$인 현상이 자주 발생하는가?
      - $\mathrm{P}(X \geq 20) = 1 - \frac{1}{\beta^\alpha}\int_0^{20} \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)}dx$
      - $y = \frac{x}{\beta}$로 두면 불완전 확률분포에서 $F(5;\ 2) = 0.96$이므로
      - $P(X \geq 20) = 1 - \int_0^5 \frac{ye^{-y}}{\Gamma(2)}dy = 1 - F(5;2) = 1 - 0.96 = 0.04$
        - 4%는 드문 경우를 나타내므로 효과적인 품질관리

감마분포와 지수분포

- 예제 6.20
  - $P(X \geq 20) = 1 - F(5; 2) = 0.04$

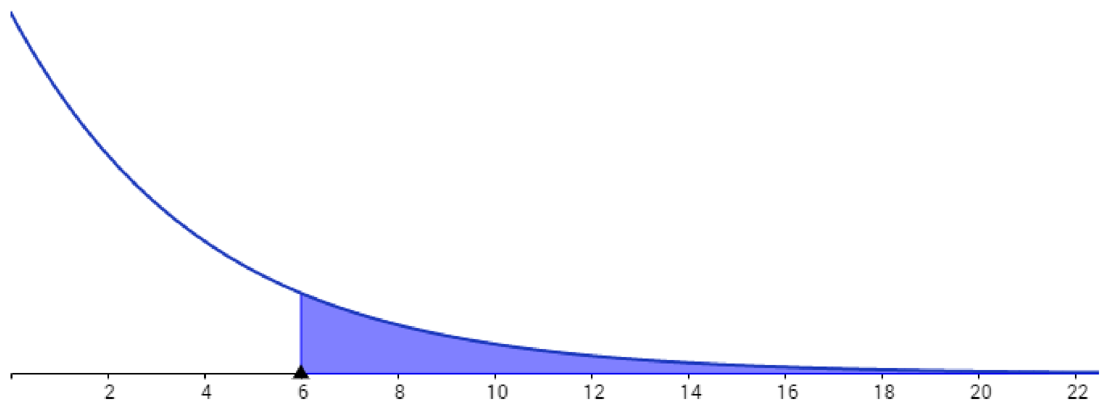


감마분포와 지수분포

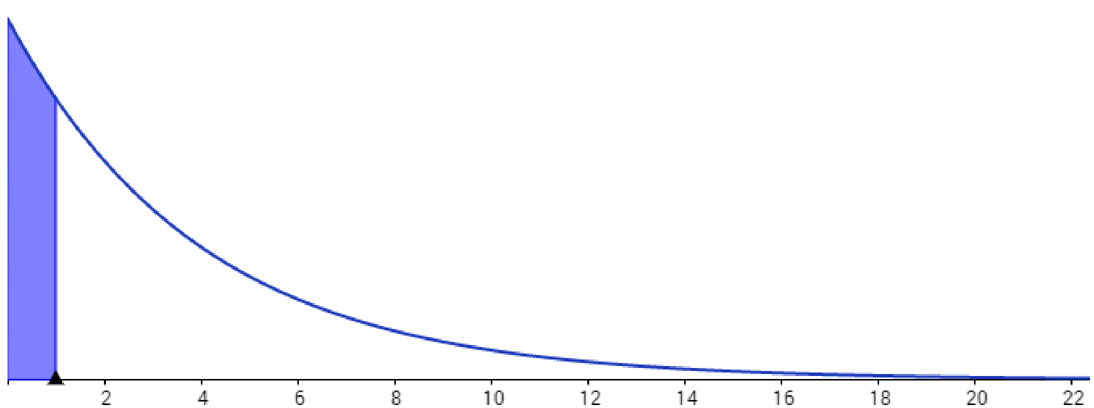
- 예제 6.21
  - 세탁기가 고장 날 때까지의 시간 Y는 다음 밀도함수를 따름
    - $f(y) = \begin{cases} \frac{1}{4} e^{-\frac{y}{4}}, y \geq 0 \\ 0, \text{구간 외} \end{cases}$
    - 해당 분포는 $\beta = 4$인 지수분포이며 지수분포의 누적분포는 다음과 같다
      - $F(y) = \frac{1}{\beta} \int_0^y e^{-\frac{t}{\beta}} dt = 1 - e^{-\frac{y}{\beta}}$
    - 6년 이상 고장 나지 않을 확률
      - $P(Y > 6) = 1 - F(6) = e^{-\frac{3}{2}} = 0.2231$
    - 1년 이내에 고장 날 확률
      - $P(Y) = 1 - e^{-\frac{1}{4}} = 1 - 0.7788 = 0.2212$

감마분포와 지수분포

- 예제 6.21
  - $P(Y > 6) = 1 - F(6) = e^{-\frac{3}{2}} = 0.2231$



  - $P(Y) = 1 - e^{-\frac{1}{4}} = 1 - 0.7788 = 0.2212$



카이제곱분포

- 의미
  - 감마분포에서 $\alpha = \frac{v}{2}, \beta = 2$인 확률분포
    - v는 양의 정수이며,
      자유도 v인 **카이제곱분포**(Chi-square Distribution)
    - 모분산의 추론에 응용
- 정의
  - 연속확률 변수 X의 확률분포가
    - $f(x;v) = \begin{cases} \frac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} x^{\frac{v}{2}-1} e^{-\frac{x}{2}}, x > 0 \\ 0, \text{구간 외} \end{cases}$ 이면, X는 자유도 v인

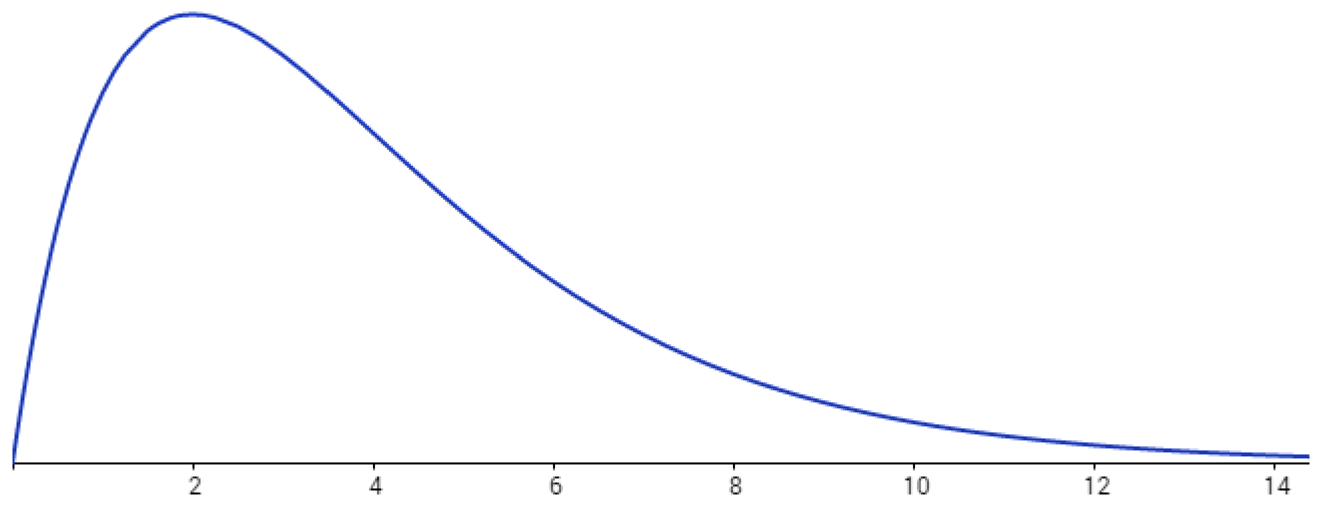
      카이제곱분포를 따름

카이제곱분포

- 특징
  - 모분산의 추론에 응용
    - 확률변수 X가 표준정규분포(Z)를 따를 때 자유도가 k인 카이제곱분포를 따름
      - $X = \sum_{i=1}^{k} Z_i^2$
    - 자유도
      - 표본 자료 중 모집단에 대한 정보를 주는 독립적인 표본 자료의 수
      - e.g., 표본분산 $s^2 = \frac{\sum(X-\bar{X})^2}{n-1}$

카이제곱분포

- 특징
  - 자유도(k=4)에 따른 카이제곱분포 그래프
    - 표본분산으로 모분산을 추정

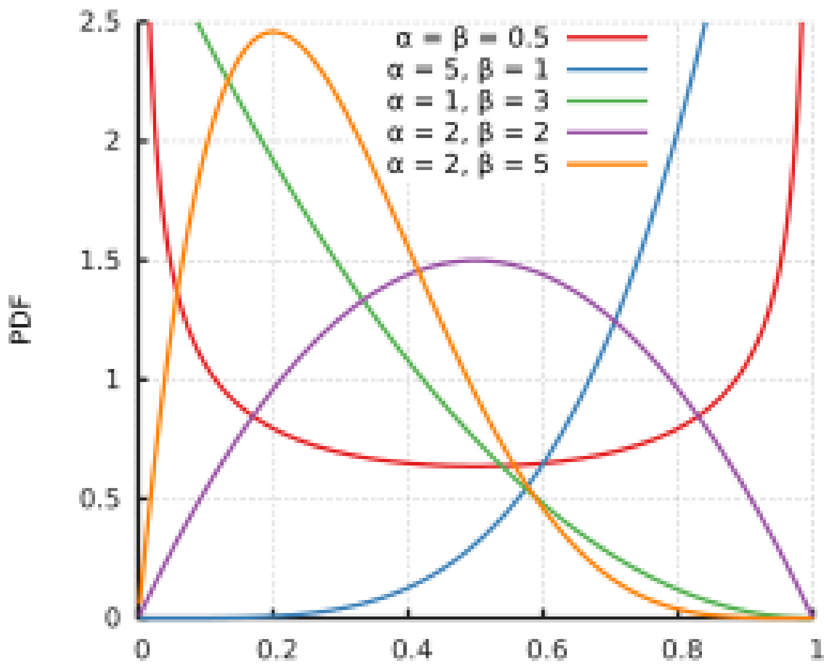


베타분포

- 의미
  - 균일분포를 일반화 한 것
  - 베타함수로 표현되며 베타함수는 다음을 따름
    - $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
    - 이항 계수(Binomial Coefficient) 즉, 조합(Combination)을 연속형 함수로 해석
- 정의
  - 연속형 확률변수 X의 밀도함수가 다음과 같을 때, 확률변수 X는 모수가 $\alpha > 0, \beta > 0$인 베타분포(Beta Distribution)
    - $f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}dx, 0 < x < 1 \\ 0, \text{구간 외} \end{cases}$

베타분포

- 특징
  - 두 모수 α, β에 따라 [0,1]구간에서 정의되는 연속확률분포
    - 모수 $\alpha = 1, \beta = 1$인 베타분포가 균일분포
    - 사전확률(Prior Probability)에 응용
      - 사전확률: 관측 하기 전 추측된 확률분포



베타분포

- 특징
  - 베타분포의 평균과 분산
    - $\mu = \frac{\alpha}{\alpha+\beta}$
    - $\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

베타분포

- 활용

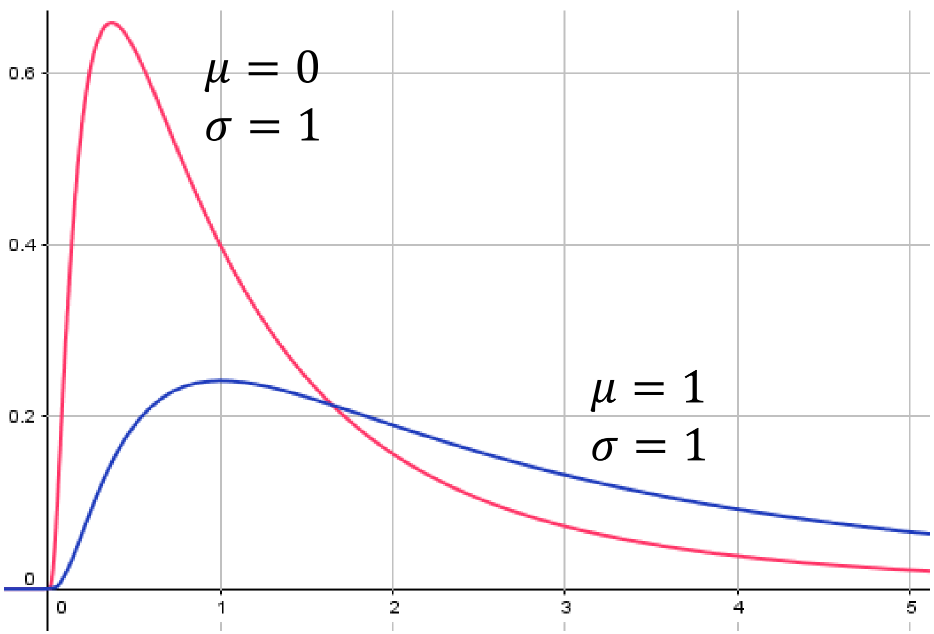
로그정규분포

- 의미
  - 정규분포 확률변수(Y)에 밑이 e인 지수를 취한 확률변수 ($X = e^Y$)가 로그정규분포를 따름

- 정의
  - 확률변수 $Y = \ln(X)$가 평균이 μ이고, 표준편차가 σ인 정규분포를 따를 때, 확률변수 X의 분포를 로그정규분포(Log-normal Distribution)라고 하며, X의 밀도함수는 다음을 따름
    - $f(x; \mu, \sigma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}[ln(x)-\mu]^2}, x \geq 0 \\ 0, x < 0 \end{cases}$

로그정규분포

- 특징
  - 로그정규분포의 평균과 분산
  - $\mu = e^{\mu+\frac{\sigma^2}{2}}, \sigma^2 = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right)$
- 로그정규분포 그래프

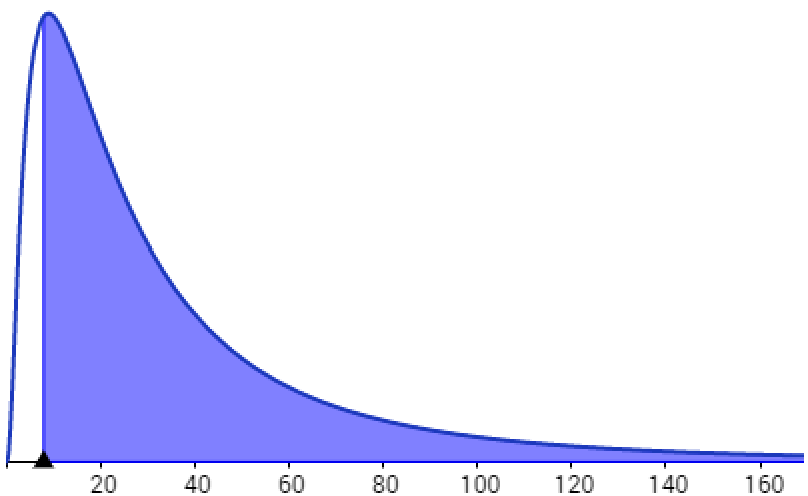


로그정규분포

- 예제 6.22
  - 화학공장에서 배출되는 오염물질의 농도(ppm)가 $\mu = 3.2$이고 $\sigma = 1$인 로그정규분포를 따를 때, 농도가 8ppm을 초과할 확률
    - 오염물질의 농도를 X라 하면
      - $P(X > 8) = 1 - p(X \leq 8)$
    - $\ln(X)$는 $\mu = 3.2$이고 $\sigma = 1$인 정규분포를 따르므로,
      - $P(X \leq 8) = \Phi\left[\frac{\ln(8) - 3.2}{1}\right] = \Phi(-1.12) = 0.1314$
        - Φ=표준정규분포의 누적분포함수(CDF)
      - $P(X > 8) = 1 - 0.1314 = 0.8686$

로그정규분포

- 예제 6.22
  - $P(X > 8) = 1 - 0.1314 = 0.8686$



로그정규분포

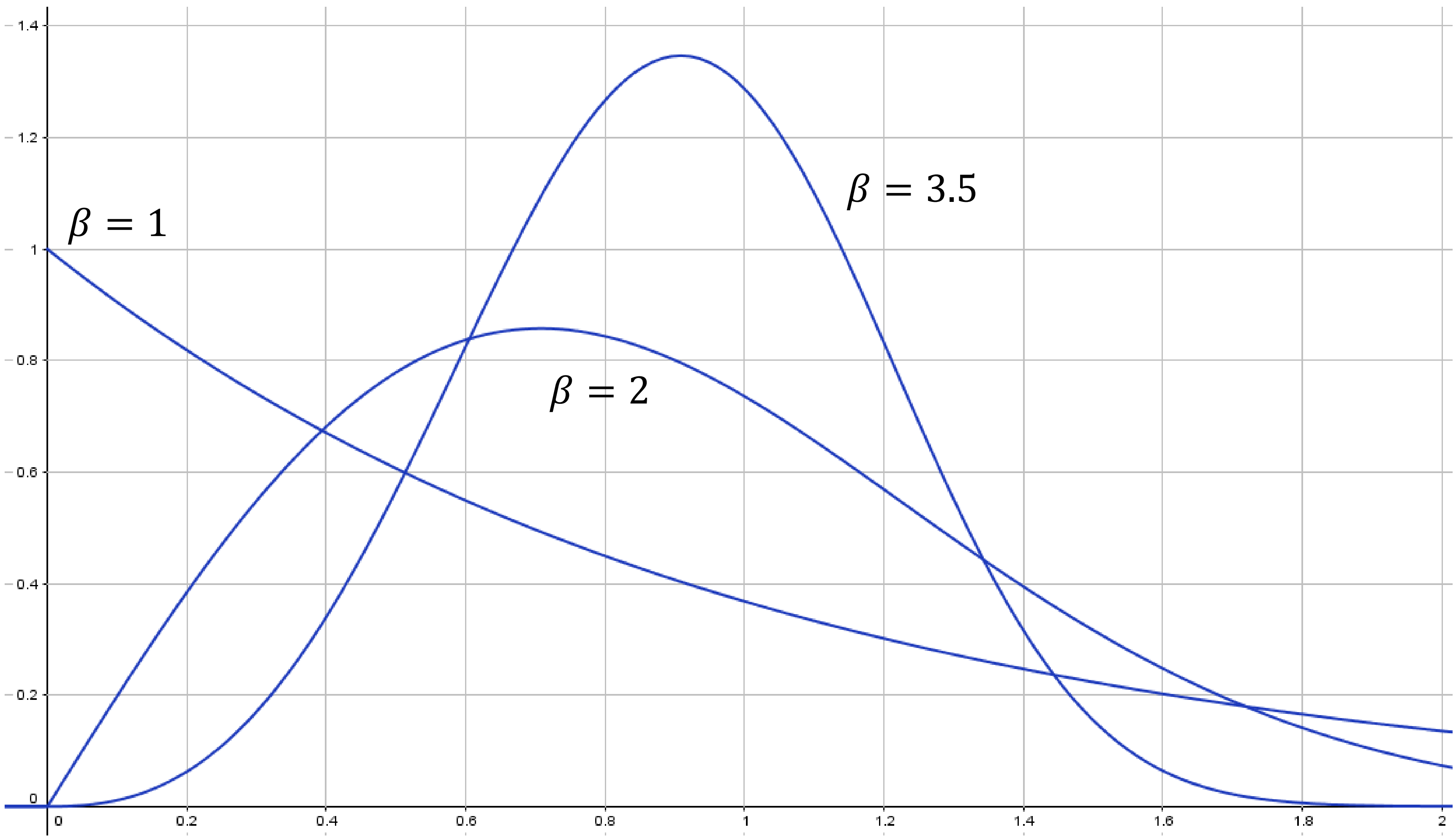
- 예제 6.23
  - 기관차에 사용되는 전자제어장치의 수명(단위: 1000마일)은 $\mu = 5.149$ 이고, $\sigma = 0.737$인 로그정규분포를 따를 때 수명의 5백분위수(=5% 에 위치하는 값)는?
    - 표 A.3에서 $P(Z < -1.645) = 0.0500$이며, 장치의 수명을 X 라고 하면 $\ln(x)$는 평균 5.149, 표준편차 0.737인 정규분포를 따름, X의 5백분위수는
      - $\ln(x) = 5.149 + (0.737)(-1.645) = 3.937$
    - 따라서 x=51.265 이며, 이는 장치의 5%만이 51,265마일 이하의 수명을 가짐

와이블분포

- 의미
  - 동일한 환경조건에 적용되는 동일한 요소(제품)이라도, 문제발생 값(시간)이 요소마다 다르고 예측할 수 없음
    - 감마분포나 지수분포를 적용할 수 있지만, 와이블분포를 적용하는 것이 바람직
- 정의
  - 연속확률변수 X의 확률분포가 다음을 따를 때, X는 모수 α, β를 갖는 **와이블분포**
    - $f(x;\alpha,\beta) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}}, x > 0 \\ 0, \text{구간 외 (단}, \alpha > 0, \beta > 0) \end{cases}$

와이블분포

- 특징
  - β값에 따른 그래프



와이블분포

- 특징
  - 와이블분포의 확률변수
    - 고장까지의 시간, 수명 등
  - 와이블분포의 평균과 분산
    - $\mu = \alpha^{-\frac{1}{\beta}}\Gamma\left(1 + \frac{1}{\beta}\right)$
    - $\sigma^2 = \alpha^{-\frac{2}{\beta}}\left\{\Gamma\left(1 + \frac{2}{\beta}\right) - \left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2\right\}$
  - 와이블분포의 누적분포함수
    - $F(x) = 1 - e^{-\alpha x^{\beta}}, x \geq 0, \alpha > 0, \beta > 0$

와이블분포

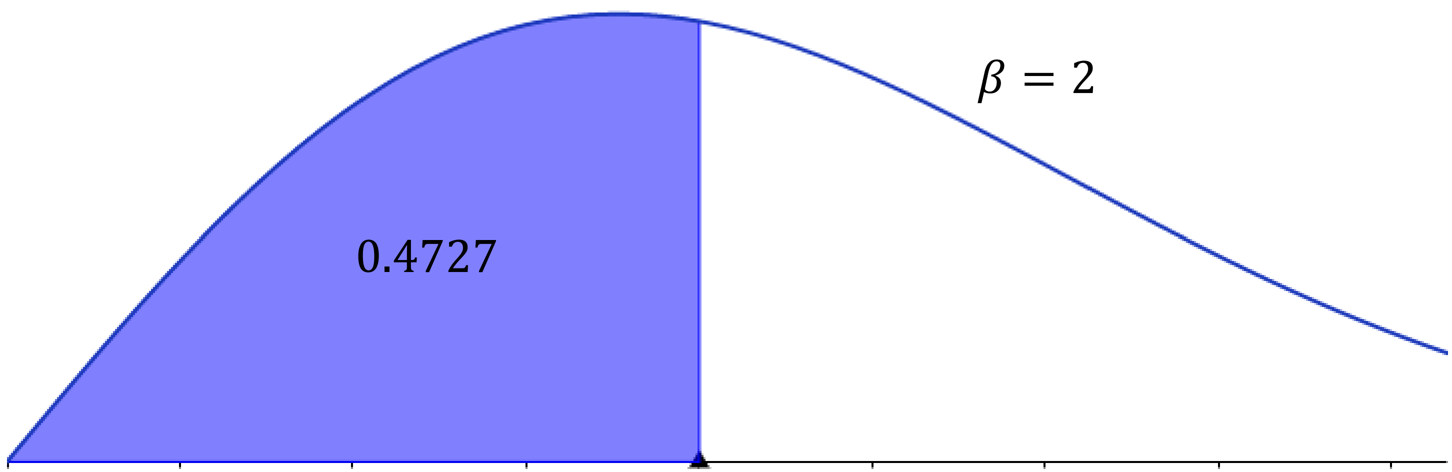
- 특징
  - 고장률(Failure Rate or Hazard Rate)
    - $R(t)$를 시점 t에서 주어진 부품의 신뢰도로 정의하면
      - $R(t) = P(T > t) = \int_t^\infty f(t)dt = 1 - F(t)$
    - $F(t)$는 T의 누적분포이며, 부품이 시간 t까지 작동했을 때, 구간 $T = t$와 $T = t + \Delta t$사이에 고장 날 조건부확률
      - $\frac{F(t+\Delta t)-F(t)}{R(t)}$
    - Δt로 나누고 $\Delta t \rightarrow 0$으로 극한을 취하면, $Z(t)$로 표시되는 **고장률**을 유도
      - $Z(t) = \lim_{\Delta t \rightarrow 0} \frac{F(t+\Delta t)-F(t)}{\Delta t} \frac{1}{R(t)} = \frac{F'(t)}{R(t)} = \frac{f(t)}{R(t)} = \frac{f(t)}{1-F(t)}$

와이블분포

- 특징
  - 고장률(Failure Rate or Hazard Rate)
    - 정리하면, 와이블분포의 시간 t에서 **고장률**
      - $Z(t) = \alpha\beta t^{\beta-1}, t > 0$
    - 의미
      - 고장률은 시간 t까지 고장 나지 않았다고 할 때, t부터 Δt만큼까지 변화율
        - 변화율이 증가하는지 감소하는지 여부가 중요
        - $\beta = 1$이면 고장율은 상수α이며, 지수분포의 건망성 특성을 나타냄
        - $\beta > 1$이면 $Z(t)$는 증가함수이며, 시간이 지남에 따라 마모됨을 나타냄
        - $\beta < 1$이면 $Z(t)$는 감소함수이며, 시간이 지남에 따라 강해짐을 나타냄

와이블분포

- 예제 6.24
  - 어느 제품의 수명(단위: 시간)은 $\alpha = 0.01, \beta = 2$인 와이블 분포를 따를 때, 8시간 이전에 이 제품이 고장 날 확률
    - $P(X < 8) = F(8) = 1 - e^{-(0.01)8^2} = 1 - 0.5273 = 0.4727$



  - $\beta = 2$이므로 시간에 지남에 따라 마모됨
    - $Z(t) = 0.02t$로 주어짐

감사합니다!