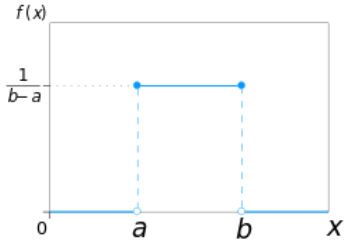
Probability Distribution

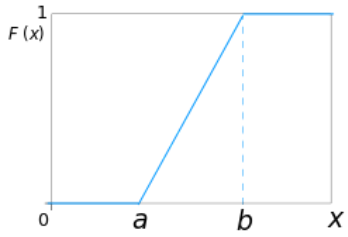
연속균등분포 (Continuous Uniform Distribution) $U(a, b)$

PDF



$$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

CDF



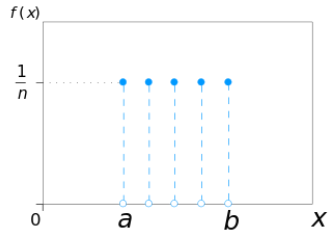
$$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b] \\ 1 & \text{for } x > b \end{cases}$$

: 분포가 특정 구간 내에서 균등하게 나타나는 경우

- 기댓값: $\frac{a+b}{2}$, 분산: $\frac{(b-a)^2}{12}$

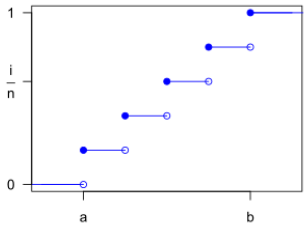
이산균등분포 (Discrete Uniform Distribution) $U(a, b)$

PDF



$$\frac{1}{N}$$

CDF

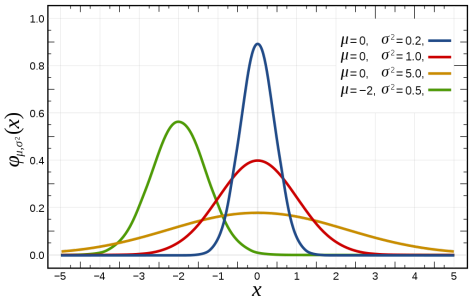


$$\frac{\lfloor k \rfloor - a + 1}{n}$$

: 함수가 정의된 모든 곳에서 값이 일정한 분포

Ex) 모든 면의 나올 확률이 동등한 주사위

정규분포 (Gaussian/Normal Distribution)



$$N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- 연속확률분포
- 기댓값 (μ)과 표준편차(σ)만으로 정의
  - $\mu = 0, \sigma = 1$이면 **표준 정규 분포** (붉은 곡선)
- 곡선의 특성
  - 종 모양, 기댓값을 중심으로 대칭형. 분산(σ^2)이 클수록 종의 폭이 넓어짐
  - 곡선과 x축 사이의 넓이=1
- **중심극한정리**: 독립 확률 변수들의 평균의 분포는 정규분포에 가까워지는 성질이 있음
- 여러 사회적, 생물학적, 자연적 현상을 모델링하는 데 자주 사용됨

베르누이 분포 (Bernoulli Distribution)

$$\text{Bern}(x; p) = \begin{cases} p & \text{if } x = 1, \\ 1-p & \text{if } x = 0 \end{cases}$$

- **베르누이 시행**: 결과가 두 가지 중 하나로만 나오는 실험
- **베르누이 확률변수**: 시행의 결과에 따라 성공이면 1, 실패면 0의 값을 갖는 확률 변수 X
- 이산확률분포
- 각 시행이 성공일 확률 $= p$, 실패일 확률 $= 1 - p = q$
- 기댓값 $= p$, 분산 $= pq$

```
# Ex) [0,1]에서 1이 나오는 횟수
p = 0.6 # 1이 나올 확률
rv = scipy.stats.bernoulli(p) #확률변수
rv.pmf([0, 1]) #array([0.4, 0.6])
```

이항분포 (Binomial Distribution)

$$Bin(x; n, p) = \binom{n}{k} p^x (1-p)^{1-x}, \quad x = 0, 1, ..., n$$

: **베르누이 시행**을 n번 시행하는 이산확률분포

- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$는 **이항계수**
- n이 충분히 크면 (np와 $\sqrt{npq}$가 5보다 클 때) 정규분포 $N(np, npq)$에 가까워짐
- 기댓값 $= np$, 분산 $= npq$

```
# Ex) 동전을 5번 던져서 앞면이 나온 횟수
N = 5
p = 0.6 # 앞면이 나올 확률
rv = scipy.stats.binom(N, p)
rv.pmf(np.arange(6))
```

다항분포 (Multinomial Distribution)

$$\mathrm{Mu}(x;N,p) = \binom{N}{x}\prod_{k=1}^{K} p_k^{x_k} = \binom{N}{x_1,\cdots,x_K}\prod_{k=1}^{K} p_k^{x_k}$$

: 이항분포의 일반화; 세 가지 이상의 결과를 가지는 사건을 반복 시행할 시 발생하는 이산확률분포

- 서로 독립적인 사건이 일어날 확률이 각각
- 차원(k)이 2인 경우 **이항분포**
- 기댓값 $= np$, 분산 $= npq$

```
# Ex) 주사위의 각 눈이 나온 횟수
N = 30
p = [0.1, 0.1, 0.1, 0.1, 0.3, 0.3] # 각 눈이 나올 확률
rv = sp.stats.multinomial(N, p) # 다항분포의 확률변수 생성
X = rv.rvs(100) # 100개의 랜덤 표본 생성
```

카테고리 확률분포 (Categorical Distribution)

$$\mathrm{Cat}(x;p) = \begin{cases} p_1 & \text{if } x = (1,0,0,\cdots,0) \\ p_2 & \text{if } x = (0,1,0,\cdots,0) \\ p_3 & \text{if } x = (0,0,1,\cdots,0) \\ \vdots & \vdots \\ p_K & \text{if } x = (0,0,0,\cdots,1) \end{cases}$$

출력벡터: $x = (x_1, x_2, ..., x_K)$

모수벡터 ("성공확률"): $p = (p_1, ..., p_K)$

: 1부터 K까지의 카테고리 or 클래스 중 한 개의 정수값을 가지는 확률변수 (=**카테고리 확률변수**)가 따르는 이산확률분포. k가 6인 카테고리 시행 = 주사위

- **원핫 인코딩 (one-hot-encoding)**: 결과를 0과 1로만 이루어진 다차원 벡터로 변형하는 것
- 기댓값$= p$, 분산$= p(1-p)$

```
p = np.array([1/6]*6)
rv = scipy.stats.multinomial(1, p)
rv.pmf(np.arange(1,7))
xx = pd.get_dummies(np.arange(1,7))
rv.pmf(xx.values)
```

감마분포 (Gamma Distribution)

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

: 0부터 무한대의 값을 가지는 모수(p)를 베이지안 방법으로 추정한 연속확률분포

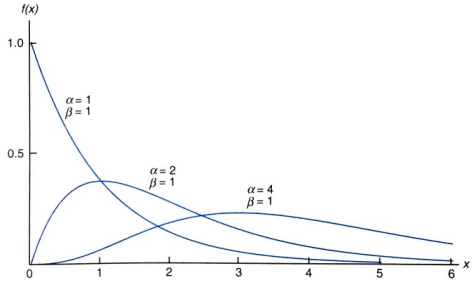
(**베이즈 추정** = 대상의 사전 확률 + 추가 정보로 사후 확률을 추론하는 방법)

: or a번째 사건이 일어날 때 까지 걸리는 시간에 대한 연속확률분포

- **감마함수**로부터 감마분포의 확률밀도함수 유도

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

- 모수
  - a=사건 개수 (형태모수; shape parameter)
  - b=사건 사이 평균 소요시간 (척도모수; scale parameter)
- 기댓값: $\frac{a}{b}$, 분산: $\frac{a}{b^2}$
- 최빈값: $\frac{a-1}{b}$
- b가 고정인 상태에서 a가 커지면 오른쪽으로 치우침




```
# Ex) a=6, b=1, 최빈값=5인 베타분포
xx = np.linspace(0, 16, 100)
scipy.stats.gamma(6).pdf(xx) # gamma 클래스에서는 b=1 고정
```

켤레 사전 분포 (Conjugate Prior Distribution)

	가능도		켤레 사전 분포
가짓수 = 2	Bernoulli A, B 둘 중 하나만 일어나는 사건	Binomial A, B가 여러 번 일어나는 사건	Beta
가짓수 > 2	Categorical A, B, ⋯ Z 중 하나만 일어나는 사건	Multinomial A, B, ⋯ Z가 여러 번 일어나는 사건	Dirichlet

: 사후분포가 사전분포와 같은 분포 계열에 속하는 경우 그 사전분포를 일컫는 말

- 사후분포의 계산이 편리해짐

베타분포 (Beta Distribution)

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

: 0부터 1까지의 값을 가지는 베르누이 분포의 모수(p)를 베이지안 방법으로 추정한 연속확률분포

- 기댓값: $\frac{a}{a+b}$, 분산: $\frac{ab}{(a+b)^2(a+b+1)}$
- 최빈값 (확률분포가 최대인 위치): $\text{mode} = \frac{a-1}{a+b-2}$
- $a = b$면 0.5를 중심으로 좌우 대칭
- 표본공간이 [0,1] ⇒ 두 개의 양수의 변수로 표현 가능

- 베이즈 통계에서 확률변수의 **사전분포** (prior distribution)로 주로 쓰임. “확률의 확률”
- $K = 2$인 **디리클레 분포**

```
# Ex) a=4, b=2, 최빈값=0.75인 베타분포
xx = np.linspace(0, 1, 1000)
scipy.stats.beta(4, 2).pdf(xx)
```

디리클레 분포 (Dirichlet Distribution)

$$\begin{aligned}\mathrm{Dir}(x;\alpha) &= \mathrm{Dir}(x_1, x_2, \cdots, x_K; \alpha_1, \alpha_2, \cdots, \alpha_K) \\ &= \frac{1}{\mathrm{B}(\alpha_1, \alpha_2, \cdots, \alpha_K)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}\end{aligned}$$

: k차원의 벡터 중 벡터의 요소가 양수이며, 모든 요소의 합이 1인 경우에 확률값이 정의되는 연속확률분포

Ex) 가위바위보 게임 추론

- 관측 사건인 가위바위보는 다항분포에 속함
- 다항분포의 사전 켤레 분포인 디리클레 분포를 사용 (상대가 가위바위보를 내는 확률 X는 디리클레 분포를 따른다고 가정)

토픽 모델링

- 문헌은 여러 개의 주제를, 주제는 여러 개의 단어를, 그리고 모든 단어는 어떤 주제에 포함된다고 가정했을 때: 알 수 없는 (latent한) 문헌별 주제분포 & 주제별 단어분포 추론 필요
- 주제를 고르는 행위와 단어를 고르는 행위는 모두 다항 분포 ⇒ 이 분포들이 디리클레 분포를 따른다고 가정

[잠재 디리클레 할당 파헤치기] 2. 디리클레 분포와 LDA (tistory.com)

포아송 분포 (Poisson Distribution)

$$Poi(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

: 시행 횟수는 많으나 시간적, 공간적으로 발생 빈도가 낮은 사건의 발생 수를 설명하는 연속확률분포

- 모수
  - λ= 단위 시간 동안 발생하는 평균 사건 수 (기댓값)
  - n=사건이 일어나는 횟수
- 전제 조건
  - 사건 발생이 서로 독립이어야 함
  - 단위 당 발생 확률은 동일
  - 작은 구간 내 사건의 동시 발생 확률은 아주 작음
- **이항분포**의 성공률이 작고 시행횟수가 클 경우에 포아송 분포에 근사

지수 분포 (Exponential Distribution)

PDF:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{where } x \geq 0 \\ 0 & \text{where } x < 0 \end{cases}$$

CDF (주로 사용됨):

$$F(a) = 1 - e^{-\lambda a}$$

: 어떤 사건의 경과된 **시간**에 대한 연속확률분포

- 포아송 분포의 변형

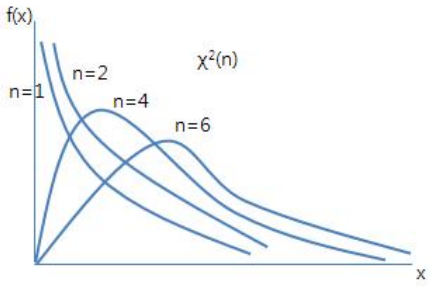
카이제곱분포 (Chi-squared Distribution)

$$\chi^2(x;\nu) = \frac{x^{(\nu/2-1)}e^{-x/2}}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}$$

: 감마 분포의 특수한 형태. "**표본정규분포**를 제곱한다"

: $X_1, X_2, ..., X_v$가 표준정규분포를 따를 때, 각각의 X를 곱하여 더한 값의 분포는 자유도가 v인 카이제곱 분포를 따른다고 정의

- v = **자유도 (degree of freedom)**

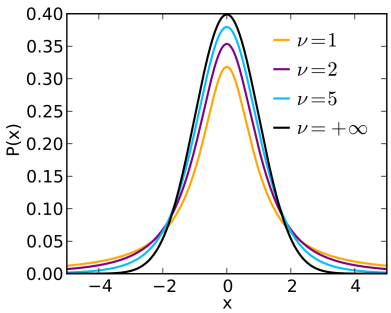


자유도 (n)에 따라 모양 달라짐

- 활용: 분산 분석
  - **카이제곱 검정 (**범주형 자료 분석; 독립성 검정, 적합도 검정)
    - 귀무가설: 관계가 독립적임
    - 대립가설: 관계가 의존적임
  - 분포 간의 차이

스튜던트 t-분포 (Student-t Distribution)

$$t(x;\mu,\lambda,\nu) = \frac{\sqrt{\lambda}}{\sqrt{\nu\pi}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(1+\lambda\frac{(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$



: 정규분포의 꼬리 부분이 두꺼워지는 팻 테일 (fat tail) 현상이 나타나는 데이터에 적합한 연속확률분포

- 모수
  - λ=정규분포의 정밀도; $(\sigma^2)^{-1}$
- 정규분포를 따르는 확률변수 N개의 표본 합 (or 평균)을 표본 분산으로 정규화하면 t분포를 따름

- 자유도에 따라 분포가 달라지기 때문에 t-분포표 참고 (함수 직접 계산 X)

확률분포 관계도