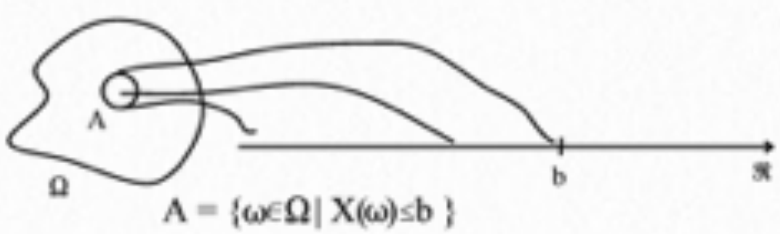
1

유명한 확률분포함수

1. 확률변수 $X(w) = x$

확률실험의 표본공간 S의 원소에 실수 값을 대응한 함수를 확률변수라 하고 X, Y, Z... 대문자 알파벳으로 표현한다.

=데이터(개체의 개별 값을 알 수 없고 변하므로)



확률실험 random experiment : 실험의 결과를 알 수 없음

원소 element : 확률실험의 개별 결과, 기호 w

표본공간 sample space : 확률실험의 모든 원소의 모임, 기호 S

확률변수 종류 : 이산형 discrete = 유한한 원소, 연속형 continuous = 무한한 원소 = 임의의 어떤 구간에도 결과(원소) 값이 존재

2. 확률분포함수 Prob. Density Function

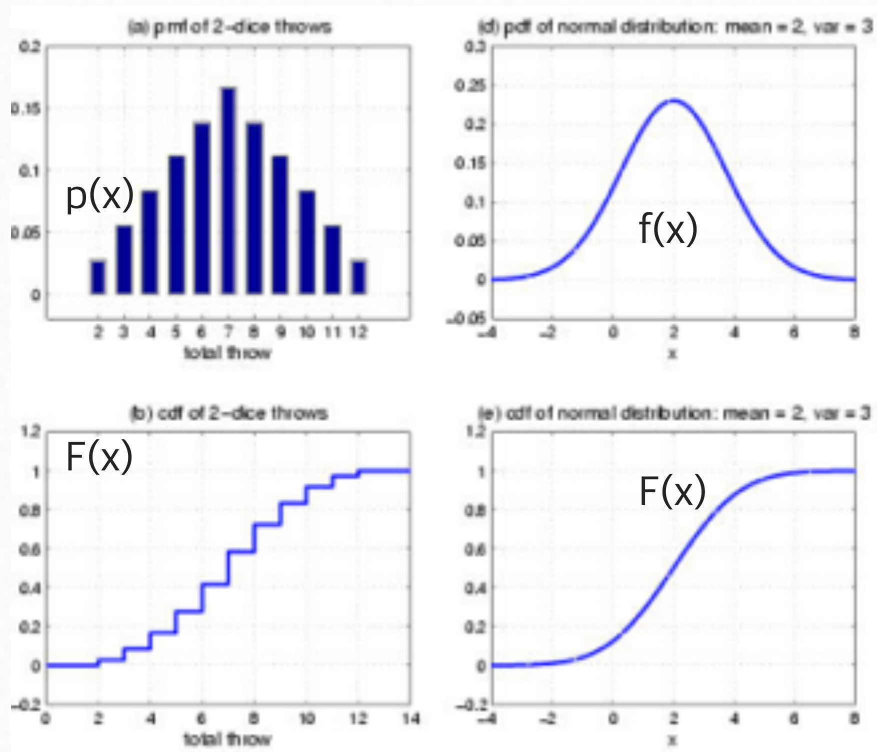
확률변수의 값이 정의역, 각 값에 대응하는 확률 값을 공역으로 하는 규칙 (함수, 표, 그래프)

(기호) $P(X = x),\ P(x),\ P_X(x) : f(x), p(x)$

확률 정의 : 각 원소 발생 가능성 동일 equally likely

누적 cumulative 확률분포함수 : $F(x) = P(X \le x)$

$F(x) = P(x) - P(x -), F'(x) = f(x)$



3. 기대값 expected

확률변수의 결과 값이 무한히 실현되었을 때 나타나는 평균적으로 기대되는 값

$$E(X) = \sum_x xp(x) = \int xf(x)dx$$

$$E(g(X)) = \sum_x g(x)p(x) = \int g(x)f(x)dx$$

4. 이산형 확률변수

1) 베르누이 분포 $X \sim B(n=1, p)$

(정의) $X =$ 베르누이 시행의 결과

(X의 범위) $X = 0, 1$

(확률밀도함수) $p(x) = p^x(1-p)^{1-x},\ x = 0, 1$

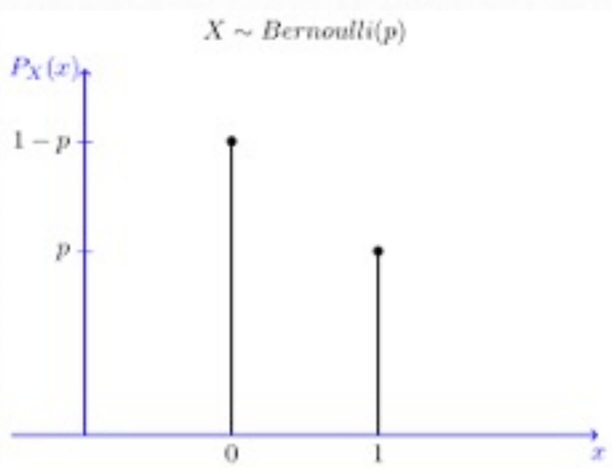
(평균, 분산) $E(X) = p, V(X) = p(1-p)$

(1) 베르누이 시행

(i) 확률실험의 결과는 이진형(binary) : 성공/실패

(ii) 성공 확률은 p로 매번 일정하다.

(iii) 모든 실험은 서로 독립이다. 각 실험의 결과는 다른 실험 결과에 영향을 미치지 않는다



2) 이항 binomial 분포 X~B(n,p)

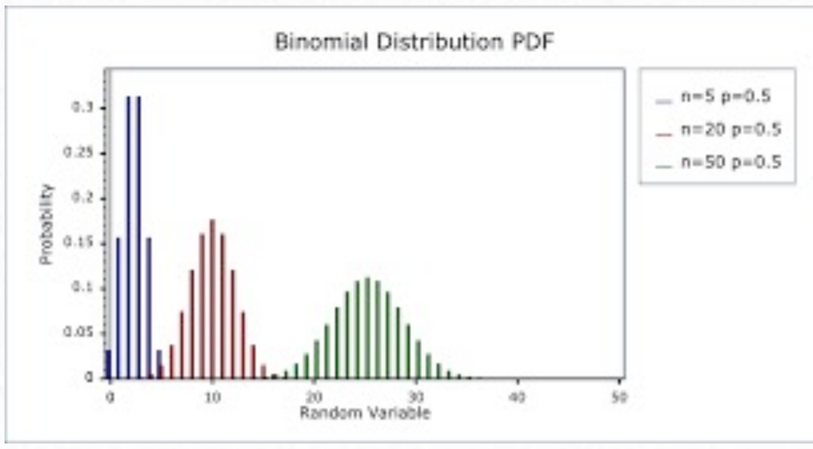
(정의) X = n번의 베르누이 시행 결과 성공의 회수

(X의 범위) $X = 0, 1, 2, \ldots, n$

(확률밀도함수)

$p(x) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0, 1, 2, \ldots, n$

(평균, 분산) $E(X) = np,\ V(X) = np(1-p)$



3) 기하 geometric 분포 X~G(p)

(정의1) X = 1번 성공까지 시행한 베르누이 회수

(정의2) 베르누이 시행에서 1번 성공까지 실패 회수

(X의 범위1) $X = 1, 2, 3, \ldots$

(X의 범위2) $X = 0, 1, 2, \ldots$

(확률밀도함수1) $p(x) = p(1-p)^{x-1},\ x = 1, 2, 3, \ldots$

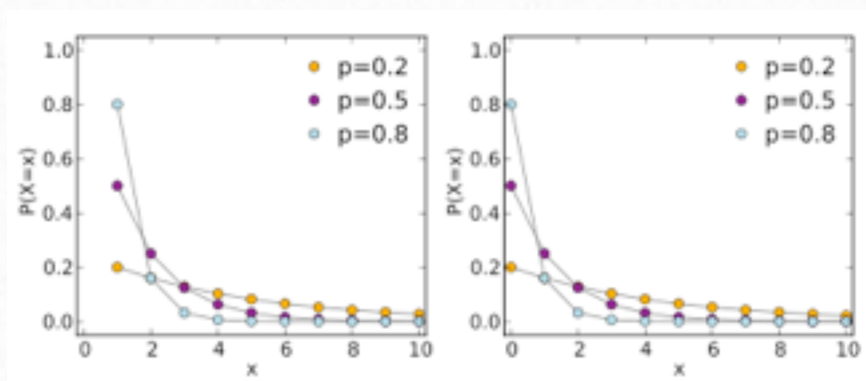
(확률밀도함수2) $p(x) = p(1-p)^x,\ x = 0, 1, 2, \ldots$

(평균1) $E(X) = 1/p$

(분산1) $V(X) = (1-p)/p^2$

(평균2) $E(X) = (1-p)/p$

(분산2) $V(X) = (1-p)/p^2$ (분산1과 동일)



성질 : 무기억성 memoryless

$P(X > x + x_0 \mid X > x_0) = P(X > x)$: 사건 발생이 x_0 시간까지 발생하지 않았다면(조건), 사건 발생이 x시간 이후 발생할 확률은 무조건 확률과 동일함

4) 음이항 Negative Binomial X~NB(r,p)

(정의) X = 베르누이 시행에서 r번 성공까지의 실패 횟수

(X의 범위) $X = 0, 1, 2, ...$

(확률밀도함수)

$$p(x) = \binom{x+r-1}{x} p^x (1-p)^r,\ x = 0, 1, 2, ...$$

(평균) $E(X) = pr/(1-p)$ (분산) $V(X) = pr/(1-p)^2$

관계

1) $NB(r=1, p) \sim Geo(p)$

2) 독립인 기하분포 합은 음이항분포

3) 음이항의 이름은 이항분포와 반대 개념에서

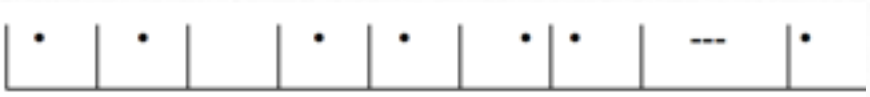
4)포아송 Poisson 분포 X~Poi(λ)

예제

- 단위 시간, 면적에서 임의의 사건 성공 회수에 관심을 갖는 경우
- 한남대 앞 정류장에 도착하는 버스 수(시간 당), 한 페이지 당 오타 숫자, 은행 창구를 찾는 고객 수(1일) 등이 예이다.

포아송 프로세스

- 시간이나 면적을 각 구간에서는 많아야 하나의 사건이 있어 나도록 동일 크기의 구간으로 나누자. (이항분포 X~(n, p))



- 각 구간에서 2개 이상 사건이 일어날 가능성은 0이며
- 각 구간의 사건 발생은 독립적이고 사건 발생 확률은 동일하고
- 구간의 사건 발생확률은 구간의 크기에 비례한다.

(정의) X= 구간에서 관심사건의 발생 회수

(X의 범위) $X = 0, 1, 2, ...$

(확률밀도함수) $np = \lambda$가 되도록 n이 충분히 크고 발생확률 p가 매우 낮음

$$\lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x} = \lim_{n\to\infty}\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$=\frac{\lambda^x}{x!}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\frac{n(n-1)\dots(n-x+1)}{n^x}\left(1-\frac{\lambda}{n}\right)^{-x}$$

$$=\frac{\lambda^x}{x!}e^{-\lambda}$$

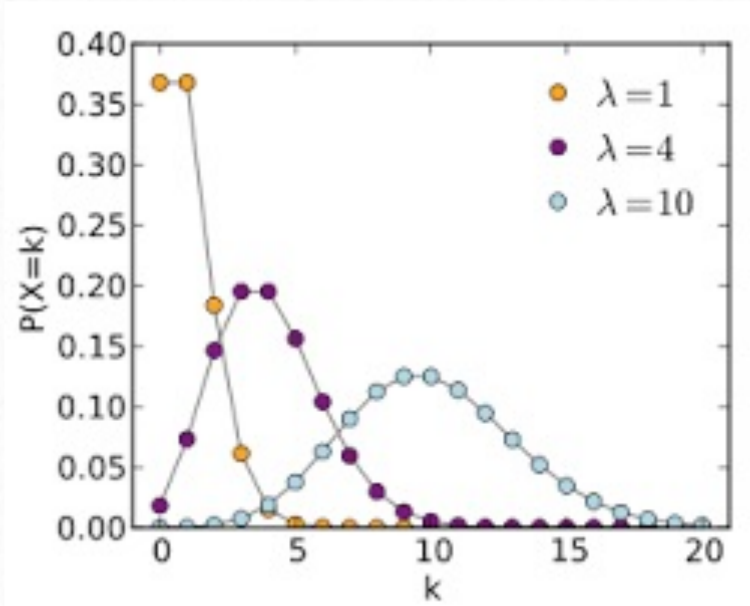
$$p(x) = \frac{\lambda^x}{x!}e^{-\lambda x},\ x = 0, 1, 2, ...$$

(평균) $E(X) = \lambda$ (분산) $V(X) = \lambda$

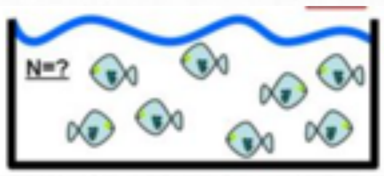
성질

1) 단위 시간 당 평균이 λ인 포아송분포의 경우 k 단위 시간 당 평균은 $k\lambda$이다

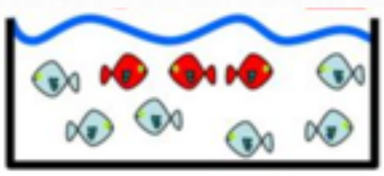
2) 독립인 포아송 분포의 합은 포아송분포이다.



6) 초기하분포 X~HG(N,K,n)



호수에 물고기가 몇 마리(N) 있을까?



색이 다른 물고기를 K 마리를 넣는다.

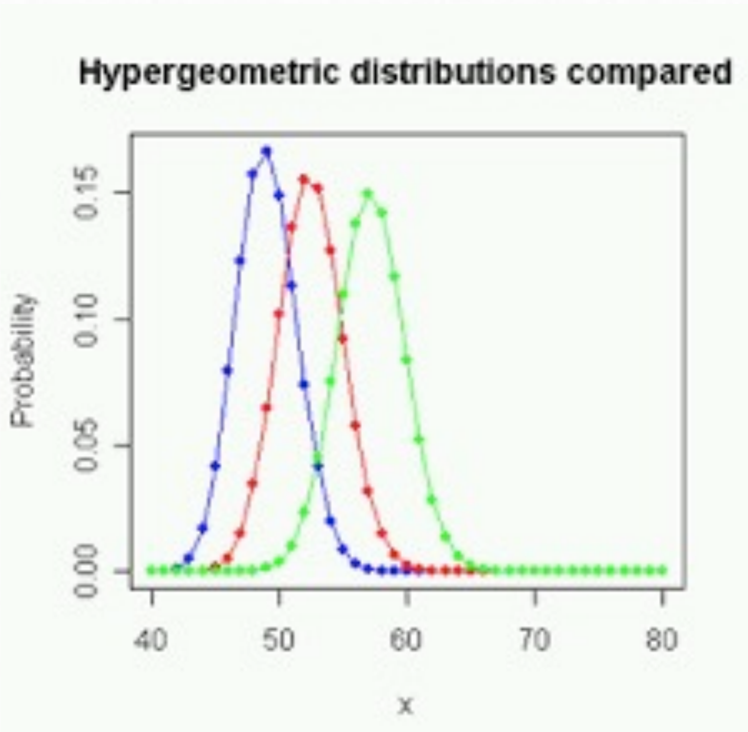
(정의) X=일정 시간이 지난 후 n마리 물고기를 잡는 실험에서 색이 다른 물고기(관심 집단) 수

(X의 범위) $X = 0, 1, 2, ..., \min(K, n)$

(확률밀도함수)

$$p(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N+K}{n}}, \ x = 0, 1, 2, ..., \min(K, n)$$

(평균) $E(X) = n\frac{K}{N}$ (분산) $V(X) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$

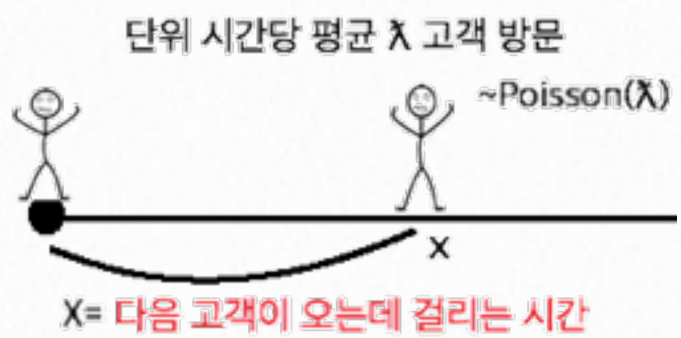


5. 연속형 확률변수

연속형 확률변수는 확률을 직접 계산할 수 있는 이산형과 달리 변수의 관계 속에서 유도됨 (정규분포만 수학적 접근방법에 의해 도출)

1) 지수분포 $X \sim Gamma(\alpha, \beta = \frac{1}{\lambda}), Exp(\beta)$

개념



단위시간 당 평균 λ명이 온다면, 고객이 온 다음 고객이 오는데 걸리는 시간은 평균적으로 $1/\lambda$이다.

포아송분포를 따르는 사건이 발생하는데 걸리는 시간을 X라고 하자.

$$F(x) = P(X \le x) = P(Y = 0 \,|\, Y \sim Poisson(\lambda x))$$

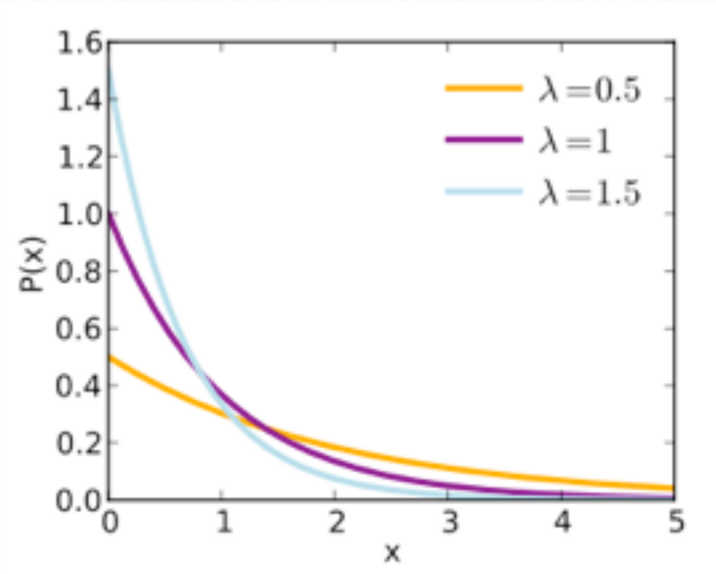
=> 누적확률밀도함수($F(x)$)의 의미는 고객이 오는데 걸리는 시간이 x보다 적을 확률은 포아송 사건이 x 이후에 일어난다는 것, 즉 x 이전에는 포아송 사건이 일어나지 않는다. $P(Y = 0 \,|\, Poisson(\lambda x))$는 λx 단위 시간 당 사건이 일어나지 않음

(정의) X = 포아송 분포를 따르는 사건이 발생하는데 걸리는 시간 (β = scale 모수, $\lambda = 1/\beta$ rate 모수

(X의 범위) $0 < x$

(확률밀도함수) $f(x) = \lambda x^{-\lambda x} = \frac{1}{\beta}e^{-\frac{x}{\beta}}, \ 0 < x$

(평균) $E(X) = \frac{1}{\lambda} = \beta$ (분산) $V(X) = \frac{1}{\lambda^2} = \beta^2$



수명에 관련된 분포에 사용, 그러나 무기억성으로 인하여 전구, 제품의 수명은 모수를 하나 더 가진 와이블(Weibull) $f(x;\alpha,\beta)=\frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1}e^{-(\frac{x}{\beta})\alpha}$,

$\alpha = shape, \beta = scale$ ($\beta=1 \sim Exp(\beta)$)

성질 : 무기억성 memoryless

$P(X > x + x_0 \mid X > x_0) = P(X > x)$

2) 감마분포 $X \sim Gamma(\alpha, \beta = \frac{1}{\lambda})$

개념

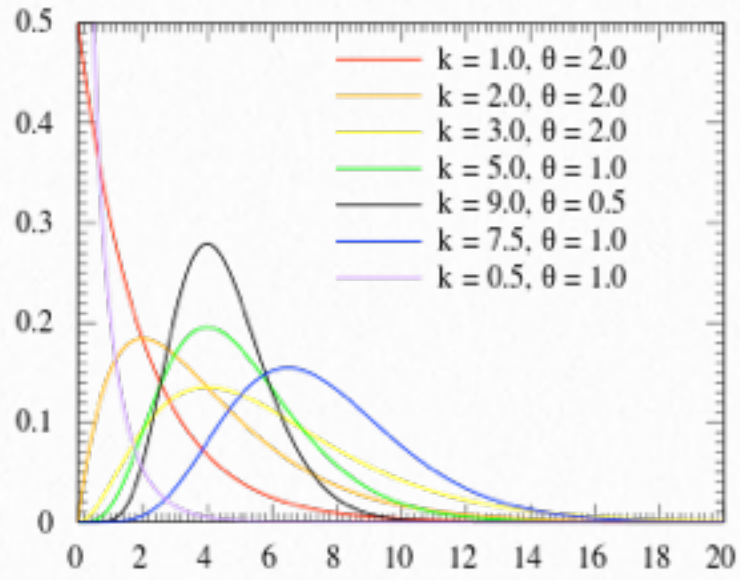
단위시간 당 평균 λ번 발생하는 포아송 사건에서, 총 α 번 일어나는데 걸리는 시간

$F(x) = P(X \le x) = P(Y \le \alpha - 1 \mid Y \sim Poisson(\lambda x))$

(정의) X = 포아송 분포를 따르는 사건이 α번 발생하는데 걸리는 시간 (α = shape 모수, β = scale 모수, $\lambda = 1/\beta$ rate 모수)

(확률밀도함수) $f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-\frac{x}{\beta}},\ 0 < x$

(평균) $E(X) = \frac{\alpha}{\beta}$ (분산) $V(X) = \frac{\alpha}{\beta^2}$



분포와 관계

1) $Gamma(\alpha = 1, \beta) \sim Exp(\beta)$

2) $Gamma(\alpha = \frac{r}{2}, \beta = 2) \sim \chi^2(r)$

3) $\frac{Gamma(\alpha_1, \beta)}{Gamma(\alpha_1, \beta) + Gamma(\alpha_2, \beta)} \sim Beta(\alpha_1, \alpha_2)$

4) 가법성: additivity 독립인 감마분포의 합은 감마분포 $X_i \sim G(\alpha, \beta) \sim (iid) \sum_{i=1}^{n} X_i \sim G(n\alpha, \beta))$

활용

모집단 분산 추론 : 카이제곱분포 : 데이터 정규분포 가정

3) 정규분포 $X \sim N(\mu, \sigma),\ Z \sim SN(0, 1)$

- 베르누이 시행의 성공의 회수는 n이 충분히 클 때 확률 근사값(이항분포의 combination 순열 값은 n이 크면 계산이 불가능, 그 시절)을 계산하기 위하여 도입되었음 (de Moivre, 1733)
- 우주 공간의 행성 간 실제 거리는 이론적 값과 오차로 이루어져 있음을 발견하고 오차에 대한 분포를 도출하게 되는데 이것이 정규분포임 (Gauss)

◉ (중심극한정리) 모집단의 분포와 상관없이 표본 크기가 충분히 큰 경우(n>20~30)표본 합, 표본평균의 샘플링 분포는 정규분포에 근사한다.

(정의) X = 오차, 좌우 대칭인 측정형 변수)

(확률밀도함수) $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$

(평균) $E(X) = \mu$ (분산) $V(X) = \sigma^2$

(표준정규분포) 평균이 0, 분산이 1인 정규분포

$f(x) = e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty$

성질

1) 표준화 $X \sim N(\mu, \sigma) \Rightarrow (\frac{X-\mu}{\sigma}) \sim SN(0,1)$

2) $Z^2 \sim \chi^2(1)$

3) 독립인 정규분포의 합 정규분포를 따른다

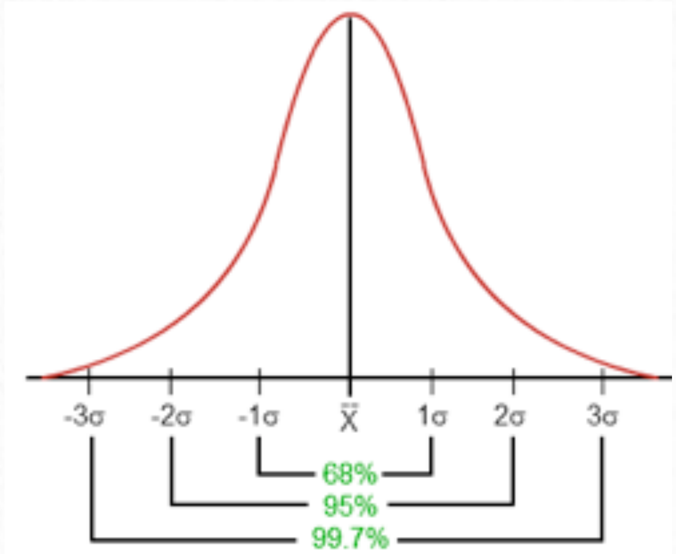
$X_i \sim N(\mu, \sigma^2) \sim (iid) \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$

활용

대표본 모집단 합, 평균 추론

활용 : 수능점수 표준화, 6-시그마 운동 :

Empirical Rule : 분포가 좌우 대칭인 경우



4) t분포 $X \sim t(df = n)$

(정의) $\frac{SN}{\sqrt{\chi^2(df=n)/n}} \sim t(n-1)$

- 데이터가 정규분포임을 가정함

(확률밀도함수)

$f(x) = complicated, \quad -\infty < x < \infty$

(평균) $E(X) = 0$ (분산) $V(X) = \frac{n}{n-2}$

성질

1) t-분포 형태는 표준정규분포와 유사함, 단 양쪽 꼬리 부분이 두꺼워 분산이 1보다 크다.

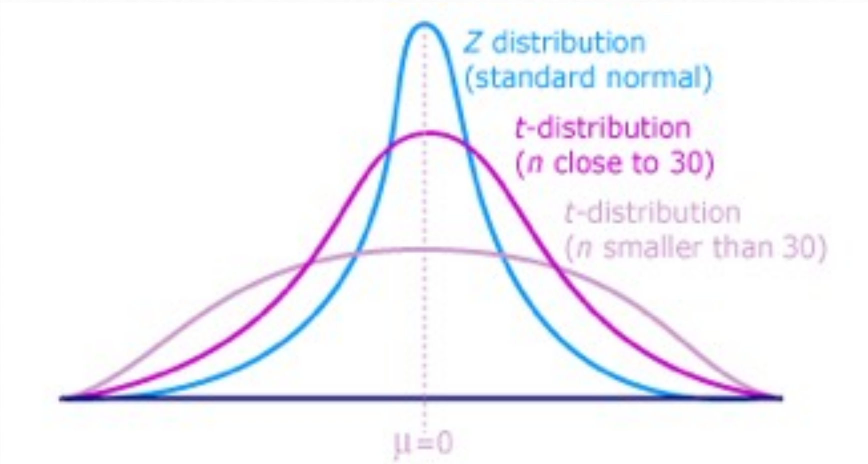
2) 단 n이 충분히 커지면 분산이 1에 가까워지고 표준정규분포에 근사한다.

3) W.S. Gosset (1908, Guinness Brewery 아일랜드) : 소표본의 경우 표본평균의 분포가 정규분포랑 다른 형태를 띠고 있음을 보고 발견한 분포

활용

소표본 모집단 평균 추론

선형모형 회귀계수 추론 (종속변수 정규분포 가정)

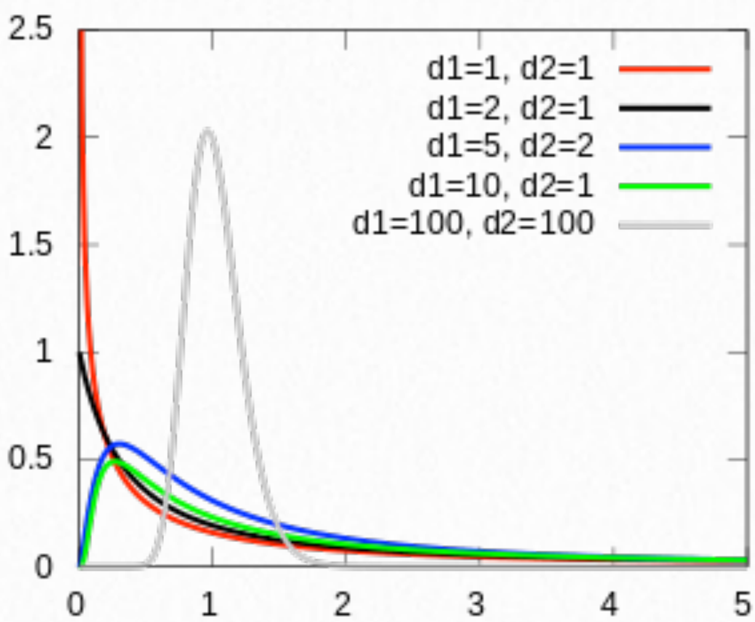


5) F 분포 $X \sim F(df_1, df_2)$

(정의) $\dfrac{\chi^2(df_1)/df_1}{\chi^2(df_2)/df_2} \sim F(df_1, df_2)$

(확률밀도함수) $f(x) = complicated,\ 0 < x < \infty$

(평균) $E(X) = \dfrac{df_2}{df_2 - 2}$ (분산) $V(X) = compicate$



활용

두 모집단 분산 차이 비교 : 데이터 정규분포 가정

분산분석 : (설명하는 변동)/(설명하지 못하는 변동)

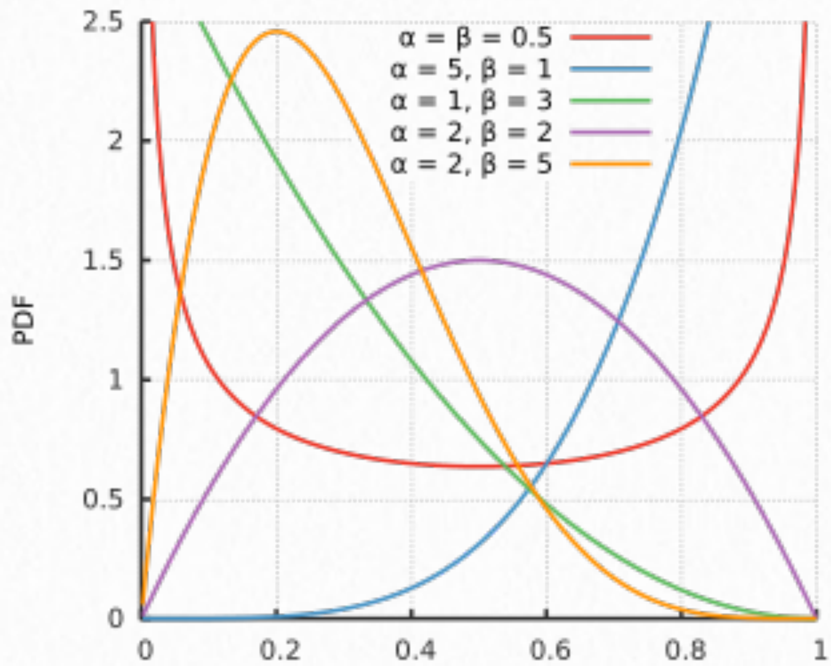
6) 베타분포 $X \sim Beta(\alpha, \beta)$

(정의) 독립인 두 감마분포의 비 $\dfrac{G(\alpha, \lambda)}{G(\beta, \lambda)} \sim Beta(\alpha, \beta)$

(확률밀도함수) $f(x) = \dfrac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$,

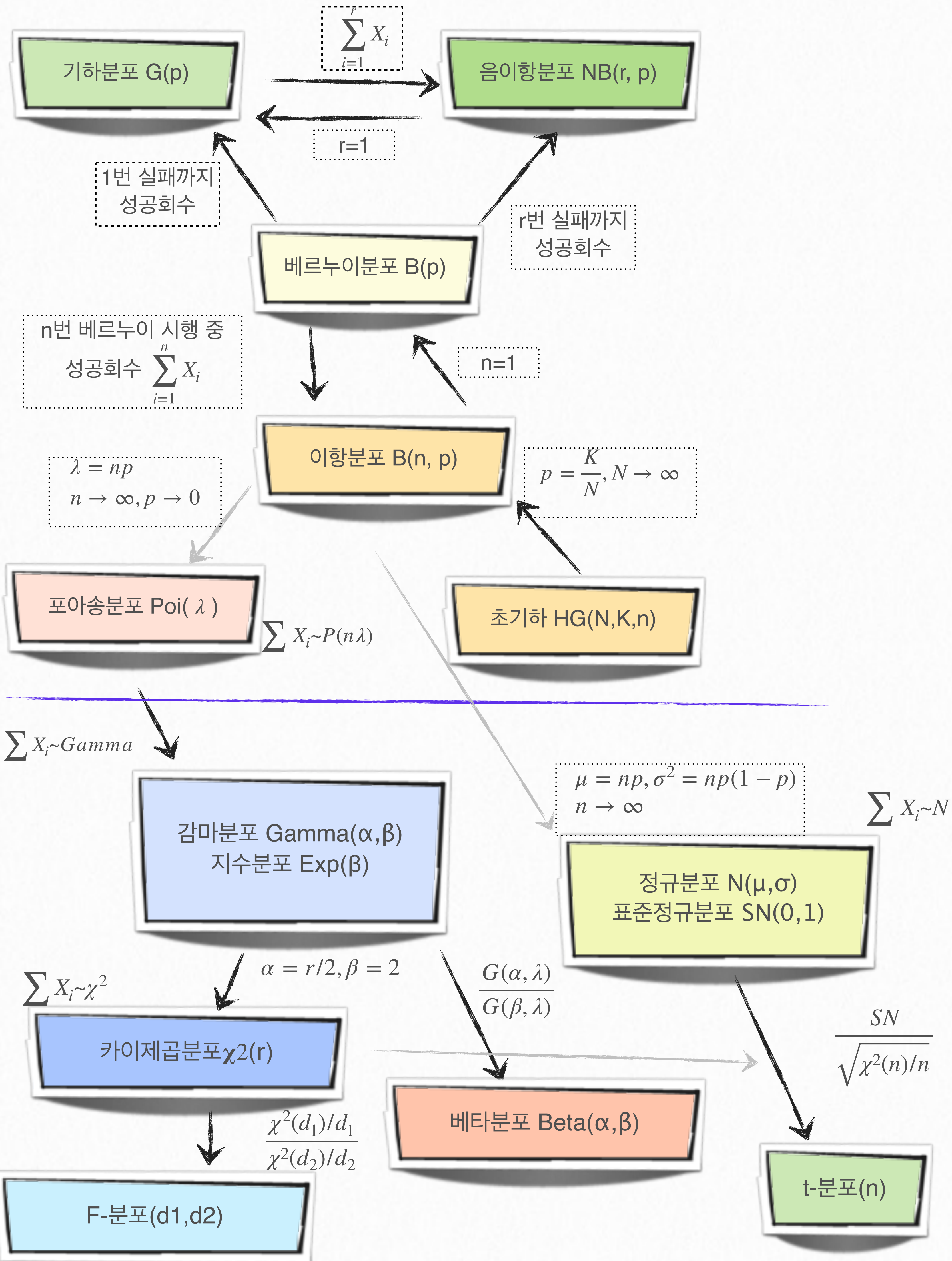
$0 < x < 1$

(평균) $E(X) = \dfrac{\alpha}{\alpha + \beta}$, 분산은 복잡



활용

베이지안 추정 시 : 모비율의 사전확률, 데이터 기븐 사후확률도 베타분포



기하분포 G(p)
$\sum_{i=1}^{r} X_i$
음이항분포 NB(r, p)
r=1
1번 실패까지 성공회수
r번 실패까지 성공회수
베르누이분포 B(p)
n번 베르누이 시행 중 성공회수 $\sum_{i=1}^{n} X_i$
n=1
이항분포 B(n, p)
$\lambda = np$
$n \to \infty, p \to 0$
$p = \frac{K}{N}, N \to \infty$
포아송분포 Poi(λ)
$\sum X_i \sim P(n\lambda)$
초기하 HG(N,K,n)
$\sum X_i \sim Gamma$
$\mu = np, \sigma^2 = np(1-p)$
$n \to \infty$
$\sum X_i \sim N$
감마분포 Gamma(α,β)
지수분포 Exp(β)
정규분포 N(μ,σ)
표준정규분포 SN(0,1)
$\sum X_i \sim \chi^2$
$\alpha = r/2, \beta = 2$
$\frac{G(\alpha, \lambda)}{G(\beta, \lambda)}$
$\frac{SN}{\sqrt{\chi^2(n)/n}}$
카이제곱분포χ2(r)
베타분포 Beta(α,β)
$\frac{\chi^2(d_1)/d_1}{\chi^2(d_2)/d_2}$
t-분포(n)
F-분포(d1,d2)