확률 probability

개념

- 확률은 미래에 발생할 사건에 대한 믿음에 대한 측정값이다.
- 확률 개념은 물리, 화학, 사회 과학 등에서 발생하는 관심 현상의 측정값을 불확실성에 의해 예측할 수 없는 경우 사용 (예) 1분간 당신 맥박 수, 다리가 무너지기 전 최대 하중 등
- 이런 상황을 랜덤(RANDOM)이라 한다. 랜덤상황이더라도 이 사건에 대한 상대 빈도(relative frequency) 정보가 있다면 예측이 가능

정의

- 확률은 관심 사건이 일어날 가능성 (chance or likelihood)을 숫자로 표현한 것
- 확률의 0 (일어날 가능성 없음)과 1(항상 일어남) 사이의 값(0% TO 100%)

확률 측정 Probability measure

상대 빈도 relative frequency

동전을 던지는 경우 {앞 면이 나올 사건}에 관심이 있어 실험을 한다고 하자. 10번을 던지니 6번이 앞 면이었다면 상대빈도는 0.60이다. 계속 100번 던지니 52번이 앞 면이 나왔다면 상대 빈도는 0.52이다. 1000번을 던지니 515번이 앞면이었다면 상대 빈도는 0.515이다.

$P(A) = \lim_{n\to\infty} \frac{f}{n}$: 관심사건 A 확률, n=실험 횟수, f=사건 A가 발생한 횟수

무한히 많은 시행 후에는 관심 사건의 나타날 가능성을 예상-확률은 무한 실험 후에 관심사건의 횟수

동전 던지기 역사

- Count Buffon (1707-1788): 4040번 동전 던지기 실험 앞면 2048회, P(앞면)= 0.5069
- Karl Pearson (1900): 24, 000 던지기 앞면 12,012, P(앞면)=0.5005
- John Kerrich : 10,000 던지기, 앞면 5067 heads, P(앞면)=0.5067.

상대 빈도 개념의 확률을 정의한다.

예 : 공정 생산 제품의 불량률(확률)에 대한 모형을 위해서는 제품 검사(확률실험)를 통하여 검사 제품 개수 중 불량품의 개수(상대 빈도)를 계산하면 된다.

Laplace 확률

표본공간의 각 원소들이 일어날 가능성이 같다고(equally likely) 가정하여 확률을 정의하는 것을 Laplace 확률(고전적 정의)이라 한다.

주사위를 던지는 실험에서 짝수가 나올 확률은 3/6=1/2으로 정의한다. 표본공간의 원소 개수는 6개이고 짝수 사건의 원소는 3개이므로 짝수가 발생할 확률은 0.50이다.

고전적 정의의 가정은 각 원소(주사위 눈금)가 나타날 확률과 동일(1/6)하다는 것이다.

일반적으로 Laplace 확률 정의는 표본공간의 개수가 유한이고 원소 모두를 알고 있을 때 사용하는데 대부분 이 정의에 의해 확률 모형이 정의된다.

확률 공리 Axiom

표본공간 S인 실험에서 임의의 사건 A에 대해 아래 조건을 만족하는 P(A)를 A의 확률(probability)이라고 정의하고 이를 확률의 공리(axiom)라 한다.

- 공리1: P(A)≥0 모든 사건의 확률 값은 0보다 크거나 같다.
- 공리2: P(S)=1 표본공간의 확률 값은 1이다.
- 공리3: 만약 임의의 두 사건 A, B 가 상호 배반적 사건이라면($A \cap B = \phi$), 합집합 확률은 $P(A \cup B) = P(A) + P(B)$로 정의된다.

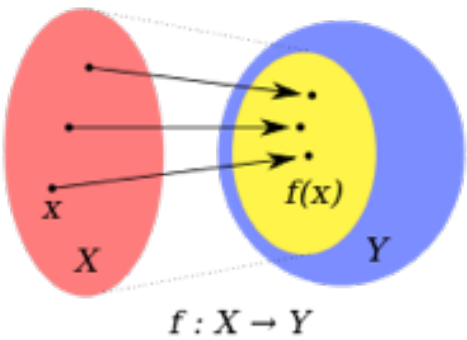
확률변수 random variable

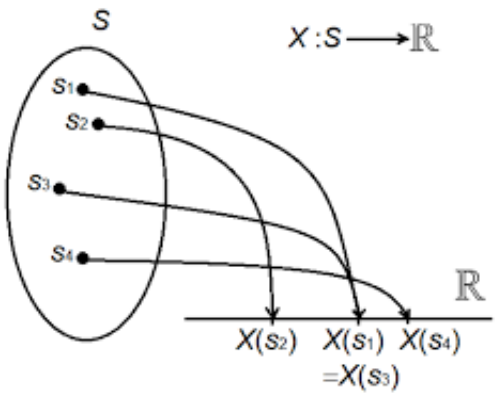
정의

확률실험 표본공간 S가 정의역(X 범위), 실수(real number)가 공역(Y 범위)인 측정 measure 함수이다.

확률실험의 각 결과에 실수 값을 할당하는 규칙 (함수) (기호) 알파벳 : X, Y, Z

함수 - $x = X(s), x = X(e), e \subseteq S$





확률변수=데이터 값과 같다.

- X=일주일 공부시간인 경우 => $x_1 = 2.5, x_2 = 10.3, x_3 = 25.0, \ldots$
- X=주사위 눈금 => $x_1 = 1, x_2 = 3, x_3 = 4, \ldots$

확률변수 종류 (1)

이산형 discrete

확률변수 X의 정의역 원소가 유한이거나 셀수 있는 경우

(예) 2개 주사위 눈금 합, 대전시 한 달 교통사고 건수

연속형 continuous

X의 정의역 원소가 무한인 경우 - 활용할 때는 구간으로 나누어 이산형처럼 사용된다.

(예) 대전시 기업 종사자 일년 연봉, 대전 월 강수량, 학생 공부시간 (분 단위로 측정하면 이산형처럼 보이지만 초 단위 이하로 측정 가능하므로 연속형이다.)

확률변수 종류 (2)

일변량 univariate

확률변수의 개수가 하나인 경우로 확률변수(데이터)의 모든 정보는 확률분포함수에 의해 얻어질 수 있다.

이변량 bivariate / 다변량 multivariate

2개 이상의 확률변수를 동시에 고려하여 관계(상관관계, 함수관계)를 분석한다. 빅데이터에서는 열(확률변수)과 행의 개수가 매우 큰 경우이다.

열과 행이 큰 경우 전통적 통계방법론을 적용하기에는 한계가 있으므로 차원을 축소하는 다양한 방법들과 시각화가 빅데이터 시대에는 중요하다.

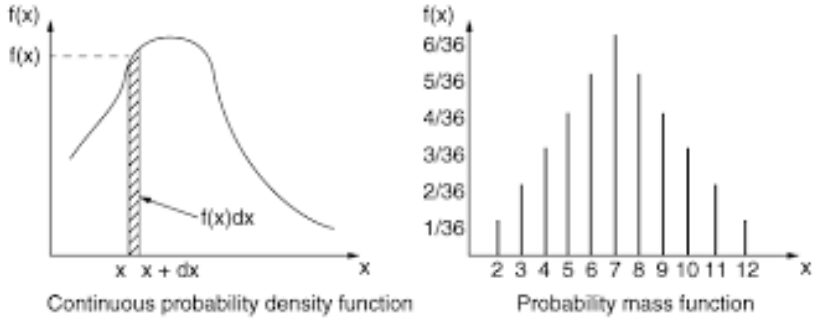
확률밀도(분포)함수 probability density/mass function

정의

확률변수의 x 값이 정의역, 각 값에 대응하는 확률값 $P(X=x)$을 공역으로 하는 규칙 : 함수, 표, 그래프 형태임

(기호) , $f(x)$ -연속형, - $p(x)$ 이산형

- 이산형 : 확률변수 각 하나의 값에 확률 대응, 막대 높이가 확률
- 연속형 : 연속형인 경우에는 각 하나의 값에에 대응하는 확률 $f(x)$은 0이다. 그러므로 연속형인 경우에는 확률은 면적

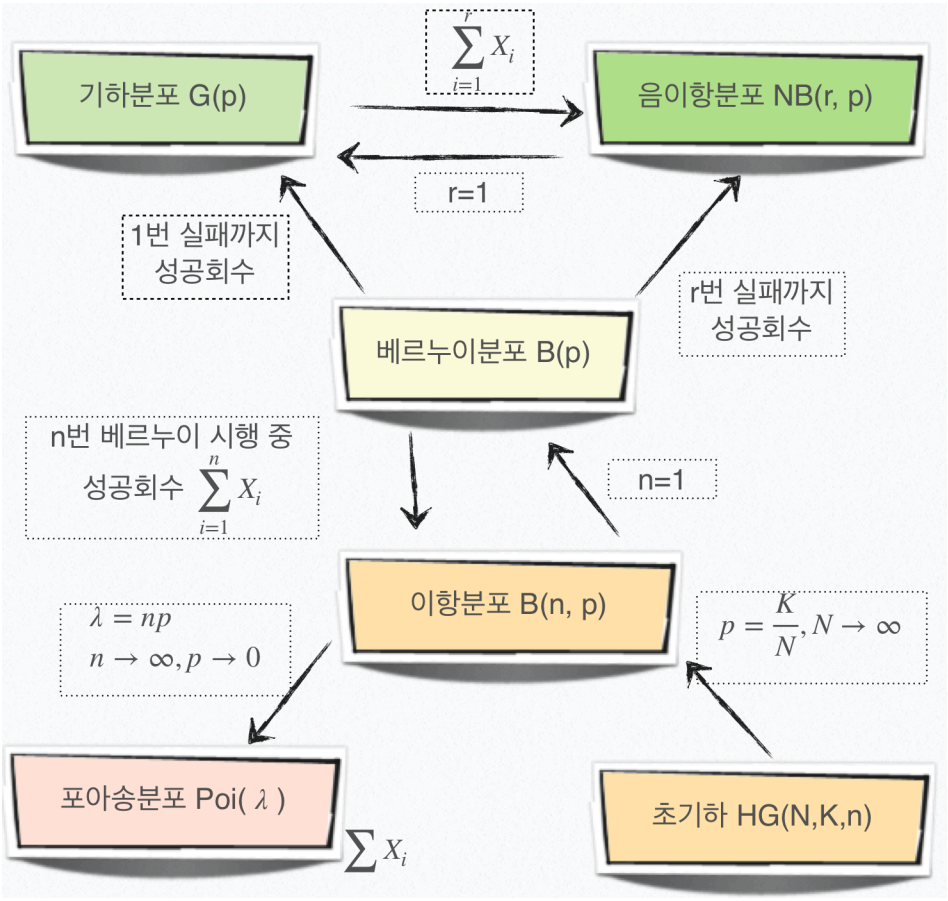


확률이므로 확률분포함수는 확률 공리를 만족한다.

- $p(x) \geq 0, f(x) \geq 0$
- $\sum_{all\ x} p(x) = 1$, $\int_{all\ x} f(x)dx$

확률분포함수는 데이터에 관한 모든 정보를 가지고 있다.

이산형 확률분포함수



베르누이 확률분포함수 $X \sim B(p)$: $f(x) = p^x(1-p)^x,\ x = 0,1$

베르누이 시행(Bernoulli experiment)

- 확률실험의 결과가 두 가지(이진형 binary dichotomous)이다. (성공 success=1, 실패 fail=0)
- 각 실험은 서로 독립(independent)이다.
- 각 실험의 성공 확률은 $p = P(X = 1)$으로 동일하다.

베르누이 확률변수 정의

확률변수 X을 베르누이 실행의 결과라 정의하자. 즉 성공이면 1, 실패면 0이다.

평균과 분산 : $E(X) = p, V(X) = pq$

이항분포 Binomial Distribution $X \sim B(n,p)$: $p(x) = \binom{n}{x} p^x(1-p)^{n-x},\ x = 0,1,2,\ldots,n$

확률분포 정의

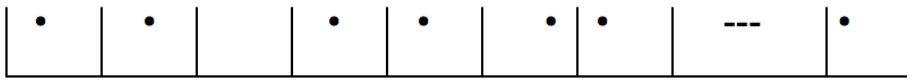
성공 확률이 p인 베르누이 시행을 n번 했을 때 성공의 회수, $X = 0,1,\ldots,n$

평균과 분산 : $E(X) = np, V(X) = np(1-p)$

포아송분포 Poisson Distribution $X \sim Poisson(\lambda)$ $p(x) = \frac{\lambda^x}{x!}e^{-\lambda}, x = 0,1,2,\ldots$

포아송 프로세스

시간이나 면적을 각 구간에서는 많아야 하나의 사건이 일어나게 동일 크기의 구간으로 나누자.

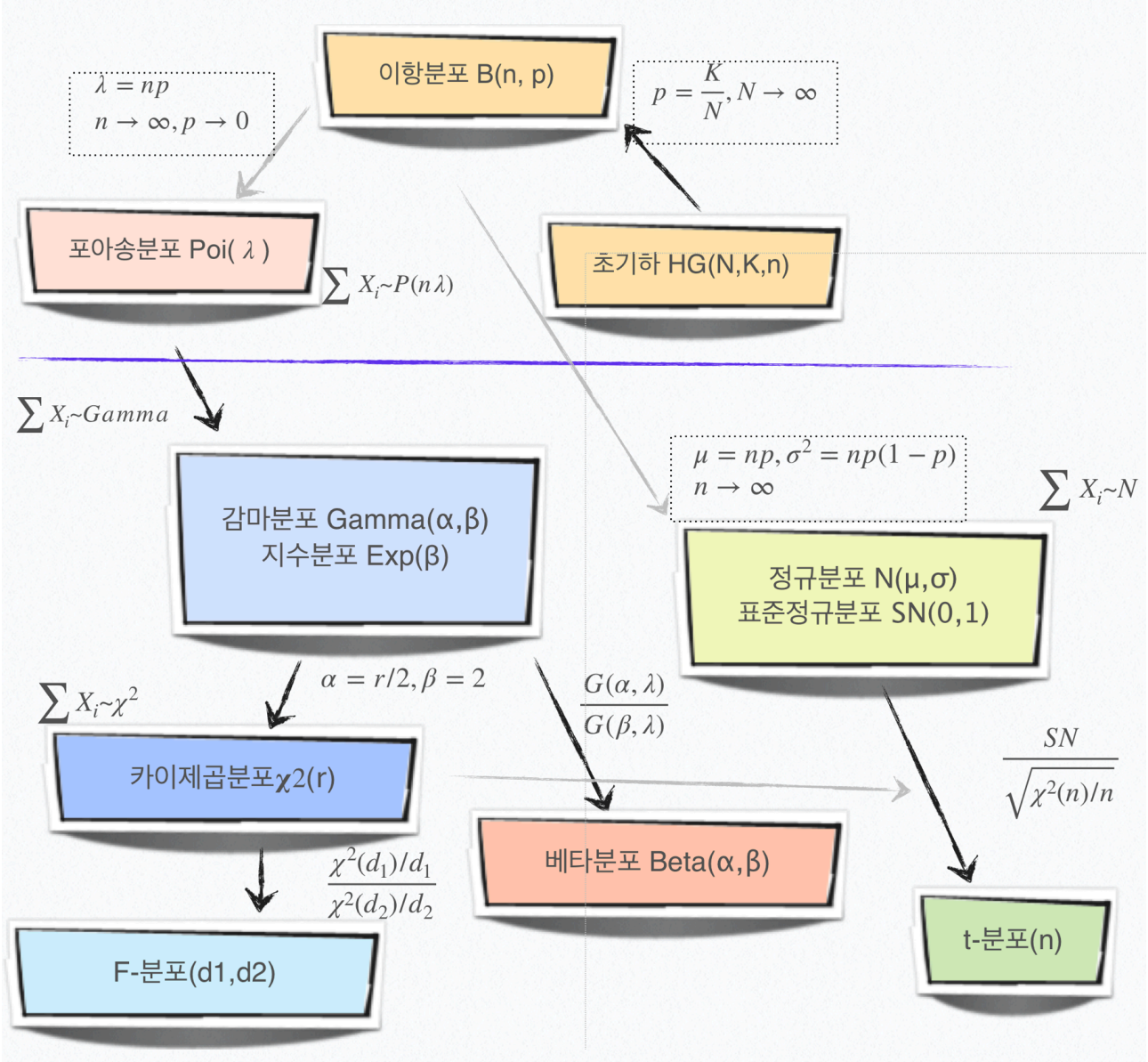


- 각 구간에서 2개 이상 사건이 일어날 가능성은 0이며각 구간의 사건 발생은 독립적이고 사건 발생 확률은 동일하고 [베르누이 시행과 동일] - 구간의 사건 발생확률은 구간의 크기에 비례한다.

확률분포 정의 및 성질

- 포아송 프로세스를 따르는 시행에서 단위 당 성공 발생 회수, $X = 0,1,2,\ldots$
- 단위 당 평균 발생 수가 λ인 관심사건의 발생 회수
- 단위구간 발생 비율은 구간의 크기에 비례한다. $X \sim P(\lambda), -> kX \sim P(k\lambda)$
- 가법성 : 독립인 두 포아송 확률변수($X_1 \sim P(\lambda_1), X_2 \sim P(\lambda_2)$) 합의 분포는 $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$

연속형 확률분포함수



균일분포 Uniform Distribution $X \sim U(a,b)$: $f(x) = \frac{1}{b-a}, \ a < x < b$

확률분포 정의

- 임의의 구간 (a,b) 어디서든 발생 가능성이 동일한 분포

균일분포 확률밀도함수 : 모수 a, b는 위치(location) 모수임

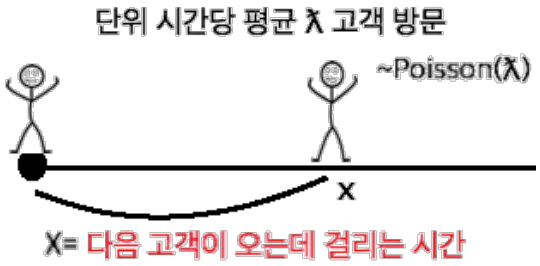
평균과 표준편차 : $E(X) = \frac{(a+b)}{2}, V(X) = \frac{(b-a)^2}{12}$

활용 : 컴퓨터에서 난수(random number)를 생성하는 분포로 이용

지수분포 Exponential Distribution $X \sim Exp(\theta = \frac{1}{\lambda})$: $f(x) = \lambda e^{-\lambda x},\ 0 < x < \infty$

확률분포 정의

- 포아송 분포를 따르는 사건들의 발생하는데 걸리는 시간
- 단위시간 당 평균 λ명이 온다면, 고객이 온 다음 고객이 오는데 걸리는 시간은 평균적으로 $1/\lambda$ 이다.



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

활용 및 성질

- 기다리는 시간의 이론 분포로 활용
- 서로 독립이며 동일 모수가 λ인 지수분포들의 합은 감마분포를 따른다.
- 무기억성 memoryless : $P(X > x + x_0 \mid X > x_0) = P(X > x)$, x_0까지 기다린 후 다시 x시간 이상 기다려야 하는 확률은 처음부터 x시간 이상 기다릴 확률과 동일하다. 즉, 지금까지 기다린 시간은 앞으로 기다리는 시간에 영향을 주지 않는다. 이로 인하여 수명에 대한 이론 분포는 와이블 분포를 사용한다.

감마분포 Gamma Distribution $X \sim Gamma(\alpha, \beta = \frac{1}{\lambda})$: $f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}},\ 0 < x$

개념

- 단위시간 당 평균 λ번 발생하는 포아송 사건에서, 총 α 번 일어나는데 걸리는 시간
- (정의) X = 포아송 분포를 따르는 사건이 α번 발생하는데 걸리는 시간 (α = shape 모수, β = scale 모수, $\lambda = 1/\beta$ rate 모수)

평균 $E(X) = \alpha\beta$, 분산 $V(X) = \alpha\beta^2$

활용 및 성질

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

- 서로 독립적으로 (병렬) 연결된 α개 부품으로 구성된 모듈이 고장나는데 걸리는 시간
- 모집단 분산 추론 : 카이제곱분포 $\sim Gamma(\alpha = 2, \beta = df/2)$
- 적합성 검정 : 데이터 분포 가정 - 카이제곱 분포

정규분포 Normal Distribution $X \sim N(\mu,\sigma^2)$, $Z \sim N(0,1)$

발견

- 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) - 오차가 이 분포 향태를 따르면 정상(normal) 그렇지 않으면 비정상 abnormal에서 나옴

중심극한정리 (Central Limit Theorem)

- 모집단의 분포와 상관없이 표본 크기가 충분히 큰 경우(n>20~30) 표본 합, 혹은 표본평균의 확률분포함수는 정규분포에 근사한다.

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

성질

- 표준화 $X \sim N(\mu,\sigma) \Rightarrow (\frac{X-\mu}{\sigma}) \sim SN(0,1)$
- $Z^2 \sim \chi^2(1)$: 서로 독립인 k개 표준정규분포의 합의 분포 $\sum Z_i \sim \chi^2(k)$
- 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

- 오차, 백색잡음
- 대표본 모집단 평균, 비율에 대한 추론
- 수능점수 표준화 : (참고) t-점수=평균이 50, 표준편차가 10인 정규분포로 변환

t-분포 $X \sim t(df = n)$

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

- 데이터가 정규분포임을 가정함

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

(성질)

- t-분포 형태는 표준정규분포와 유사함, 단 양쪽 꼬리 부분이 두꺼워 분산이 1보다 크다.
- 단 n이 충분히 커지면 분산이 1에 가까워지고 표준정규분포에 근사한다.
- W.S. Gosset (1908, Guinness Brewery 아일랜드) : 소표본의 경우 표본평균의 분포가 정규분포랑 다른 형태를 띠고 있음을 보고 발견한 분포

(활용)

- 소표본 모집단 평균 추론 - 모집단 정규분포 가정이 필요함
- 선형모형 회귀계수 추론 (종속변수 정규분포 가정)

F 분포 $X \sim F(df_1, df_2)$

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

(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)

베타분포 $X \sim Beta(\alpha, \beta)$

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

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

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

(활용)

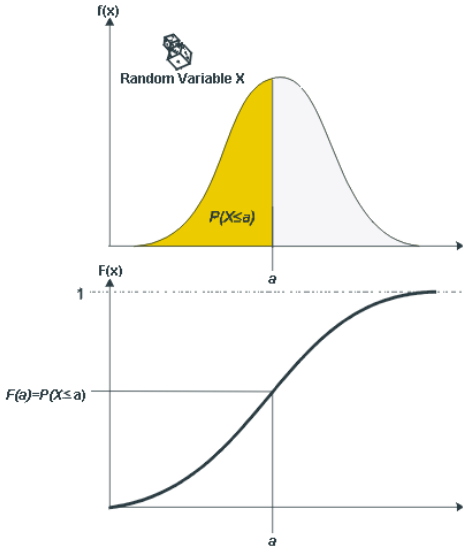
- 베이지안 추정 시 : 모비율의 사전확률, 데이터 기븐 사후확률도 베타분포

누적 cumulative 확률분포함수

정의

확률변수 X의 정의역의 가장 작은 값 $-\infty$ 부터 임의의 값 x 까지

(x 값을 포함) 확률 값을 누적시킨 함수



- (기호) $F(x) = P(X \leq x)$
- (이산형) $F(x) = \sum_{-\infty}^{x} p(x)$ (연속형)

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

성질

- 비감소 함수 non-decreasing => 만약 $x_1 < x_1$이면 $F(x_1) \leq F(x_2)$이다.
- $F(-\infty) = 0, F(\infty) = 1$

활용

확률 밀도 함수(pdf)의 특징 중 하나는 모두 적분해서 1이라는 점이다. 이는 확률의 범위가 항상 0과 1사이라는 이야기와 같다. 기본 난수 생성기의 난수 생성 범위가 0과 1사이라는 점을 이용하면 무언가 할 수 있을 것 같다.

생성하고자 하는 확률변수 X의 누적확률분포함수를 $F(x)$라 하자. $F(x)$의 역함수 $F^{-1}(x)$가 존재한다면 이를 난수생성기로 활용할 수 있다. 이를 Inverse Transformation Generating이라 한다.

만약 확률변수 U를 연속균등분포 $U(0,1)$을 따른다고 하면, 확률변수 X는 $X = F^{-1}(U)$에 의해 생성할 수 있다.

로지스틱 분포

로지스틱 분포의 누적분포함수는 $F(x) = \frac{e^x}{1+e^x}$ => $\frac{e^x}{1+e^x} = U$

이를 정리하면 $X = ln(\frac{U}{1-U}) \sim \ logistic$

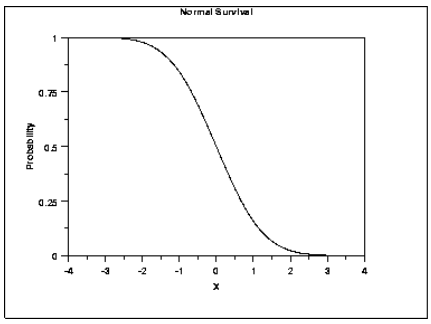
지수 분포

로지스틱 분포의 누적분포함수는 $F(x) = 1 - e^{-\lambda x}$

이를 정리하면 $X = -ln(1-U)/\lambda \sim exp(\lambda = 1/\beta)$

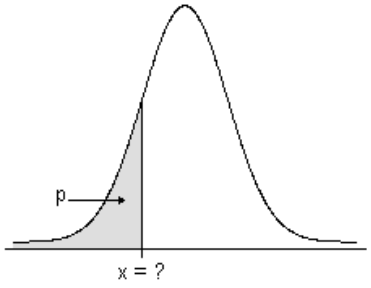
생존함수 survival function $S(x)$

$S(x) = 1 - F(x)$: 단조 감소함수 , X=환자가 사망하는 시간, 전구가 고장나는 시간



=> 정규 누적확률분포의 생존함수

Inverse CDF 역누적분포함수 $F^{-1}(x)$



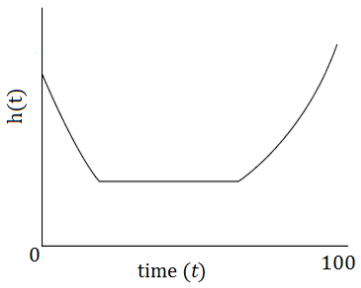
$F(x) = p => x = F^{-1}(p)$

역누적분포함수의 output은 확률변수 값이다.

위험함수 hazard function $H(x)$

$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}$, $S(x)$는 생존함수 : 시간 x까지 생존했다

가 사망(고장)할 확률



x : 사건 발생 확률변수, $f(x)$: 단위 시간당 발생 비율

$F(x)$: 시점 x 이전에 사건이 발생

$S(x)$: 사건이 시점 x 이후에 발생, 사건이 "사망"이라면 x까지 생존

위험함수는 시점 x까지 생존했다가 사건이 발생할 확률

예제 위험함수: 시간이 흐름에 따라 위험 감소, 일정, 그 후 증가함

누적 위험함수 $H(x)$: $H(x) = \int_{-\infty}^{x} h(x)dx$, $H(x) = -ln(1 - F(x))$

기대값 expected value

정의

확률변수의 결과 값이 무한히 실현되었을 때 나타나는 기대되는 값

계산) 확률변수의 각 값에 확률을 곱한 값

$$E(X) = \sum_{all\ x} p(x)(discete) \int_{all\ x} f(x)dx(continuous)$$

k-차 적률 moment

확률변수 X의 k차 승 X^k의 기대값을 k차 적률이라 한다. (정의) $E(X^k)$

평균에 대한 k-차 적률 (정의) $E(X - E(X))^k$

- 1차 적률은 평균
- 평균에 대한 2차 적률은 분산임 $V(X) = E(X - E(X))^2 = E(X^2) - (E(X))^2$

활용

(주사위 눈금의 기대값) : 주사위의 눈금을 확률변수 X라 하자.

$$E(X) = \sum_{all\ x} xp(x) = 1 * (1/6) + 2 * (1/6) + \ldots + 6 * (1/6) = 3.5$$

(데이터) 확률표본 $(x_1, x_2, \ldots, x_n)$의 평균 : 데이터 값의 중앙 척도

- 모집단 데이터 : $E(X) = \mu$, 표본 데이터 : $E(X) = \bar{x}$

(데이터) 확률표본 $(x_1, x_2, \ldots, x_n)$의 분산 : 데이터 값의 흩어짐 척도

- 모집단 데이터 : $V(X) = \sigma^2$, 표본 데이터 : $V(X) = s^2$

확률표본

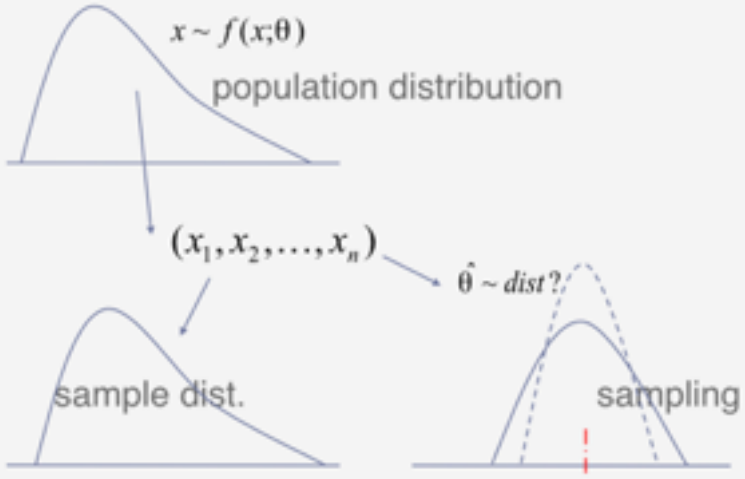
정의

확률표본은 표본을 추출할 때 모집단의 개체가 표본으로 선택될 가능성이 동일하도록 하는 경우 이를 확률표본이라 하는데, 확률표본(데이터)의 확률분포함수는 모집단의 확률분포함수와 동일하다.

$$random\ sample : (x_1, x_2, \ldots, x_n) \sim iidf(x;\theta)$$

- 확률분포는 모집단 분포와 동일한 분포이고 관측값은 서로 독립 independently and identically이다.
- 즉, 표본 데이터의 확률밀도함수는 모집단의 확률밀도함수와 동일하다.
- 그러나 연속형인 경우 히스토그램에 의존하는 방식으로는 모집단의 정확한 확률분포함수를 추정하는 것은 불가능하다.
- 확률표본으로부터 계산된 요약 값을 통계량 statistic 이라 한다. $t(x_1, x_2, \ldots, x_n)$
- 통계량은 추정에 사용되면 추정치, 검정에 사용되면 검정 통계량이라 하고 통계량의 확률분포함수를 샘플링 분포라 한다.
- 통계량의 확률밀도함수는 모집단의 분포와 상이하다. 통계량의 확률분포함수는 통계적 추론에 핵심적인 역할을 한다.
- 통계량의 확률분포함수가 알려진 통계적 방법을 모수적 parametric 방법이라 하고 모르는 경우에는 비모수적 nonparametric / distribution free 방법이라 한다.

확률분포함수 종류



모집단 분포 population (probability density) distribution $x \sim f(x)$

- 모집단 개체 관심 특성인 확률변수의 분포를 의미함
- 모집단에 대한 정보를 얻기 위하여 알아야 하는 것 : ① 확률분포함수 $f(x)$ ② 모수 θ

- 확률분포함수 $f(x)$ 는 개별 개체의 관측값의 정보이므로 관심이 적음 - 모집단 분포에 관심이 있을 경우 실시하는 분석을 "적합성 검정(Goodness of Fits Test)"이라 함. 가장 유명한 것이 정규성 검정
- 모수 θ 는 모집단 개체 특성의 요약값이므로 추론을 한다는 것은 이 값에 대한 정보를 얻는 것임

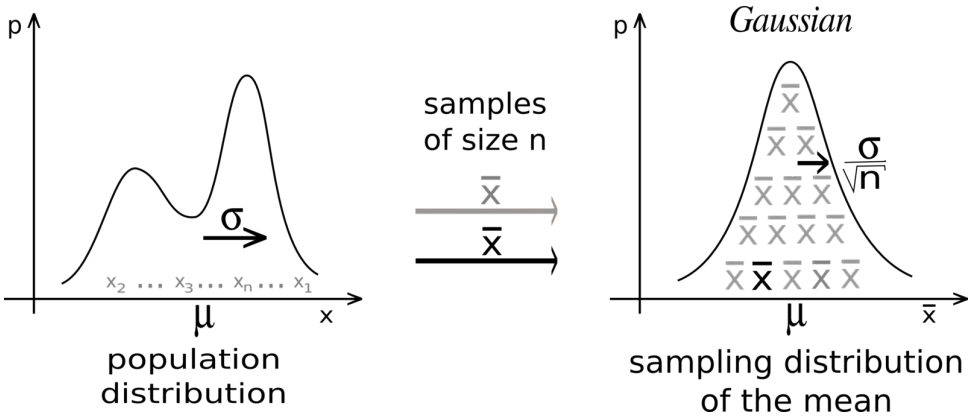
표본 분포 sample distribution $x_i \sim f(x)$

- 확률표본(iid)이므로 모집단의 분포와 동일하다. - 모집단 분포를 모르므로 표본분포도 알지 못함
- 표본분포도 표본 개체의 개별 값에 대한 정보이므로 관심의 대상이 아님

샘플링 분포 sampling dist. $t \sim f(t)$

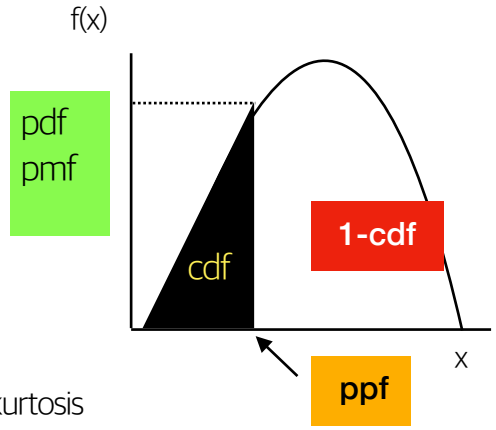
- 확률표본으로부터 계산된 통계량의 확률분포함수를 의미함 - 모집단의 분포를 알지 못해도 통계량의 분포는 알 수 있음
- 흡연여부 모집단 분포는 베르누이(p)이므로 표본조사(표본크기 n) 결과 흡연자 수는 이항분포(n, p)를 따른다. 이처럼 모집단의 분포를 알면 표본분포가 모집단의 분포와 동일하므로 통계량(흡연자 수)의 분포도 알 수 있다.
- 그러나 모집단의 분포함수를 모르면 통계량의 분포함수 샘플링분포도 알 수 없다.
- 확률표본의 히스토그램이 확률분포함수과 동일하나 연속형 데이터의 경우 히스토그램만으로는 분포함수를 추정하는 것은 불가능하다.

[중심극한 정리]
모집단의 분포와 상관없이 표본크기가 충분히 큰 경우 표본평균, 표본비율, 표본합의 확률분포함수는 정규분포를 따른다.



python 사용 Statistical functions (scipy.stats)

- rvs: Random Variates
- pdf: Probability Density Function
- cdf: Cumulative Distribution Function
- sf: Survival Function (1-CDF)
- ppf: Percent Point Function (Inverse of CDF)
- isf: Inverse Survival Function (Inverse of SF)
- stats: Return mean, variance, (Fisher's) skew, or (Fisher's) kurtosis
- moment: non-central moments of the distribution



주요 사용법

```
import scipy.stats as st
```

scipy.stats.분포함수명.rvs(모수, size=)

- 설정한 확률분포함수 따르는 데이터(난수) 생성 $generate\ x_i \sim f(x)$

scipy.stats.분포함수명.pdf (cdf, ppf) (x, arg, 모수)

- pdf - 확률분포함수 $f(x)$
- cdf - 누적분포함수 $F(x)$
- icdf - 역누적분포함수(백분위 값) $F^{-1}(x)$
- logpad, logcdf 사용 가능

scipy.stats.분포함수명.moment(n=, x, 모수)

- n차 적률 구7하기

scipy.stats.분포함수명.stats(arg, ,모수, moments='arg2')

- arg2 : m, v, s, k
- stats 대신 mean, median, var, std 사용 가능
- interval(alpha, arg, 모수) : 백분위(alpha %) 누적 백분위 값

scipy.stats.분포함수명.fit(data)

- 데이터가 설정된 확률분포를 따른다는 가정에서 모수를 추정함

이산형

bernoulli(*args, **kwds)	A Bernoulli discrete random variable.
betabinom(*args, **kwds)	A beta-binomial discrete random variable.
binom(*args, **kwds)	A binomial discrete random variable.
boltzmann(*args, **kwds)	A Boltzmann (Truncated Discrete Exponential) random variable.
geom(*args, **kwds)	A geometric discrete random variable.
hypergeom(*args, **kwds)	A hypergeometric discrete random variable.
logser(*args, **kwds)	A Logarithmic (Log-Series, Series) discrete random variable.
nbinom(*args, **kwds)	A negative binomial discrete random variable.
poisson(*args, **kwds)	A Poisson discrete random variable.
randint(*args, **kwds)	A uniform discrete random variable.

매개요인 argument

rvs(self, *args, **kwargs)	Random variates of given type.
pmf(self, k, *args, **kwds)	Probability mass function at k of the given RV.
logpmf(self, k, *args, **kwds)	Log of the probability mass function at k of the given RV.
cdf(self, k, *args, **kwds)	Cumulative distribution function of the given RV.
logcdf(self, k, *args, **kwds)	Log of the cumulative distribution function at k of the given RV.
sf(self, k, *args, **kwds)	Survival function (1 - **cdf**) at k of the given RV.
logsf(self, k, *args, **kwds)	Log of the survival function of the given RV.
ppf(self, q, *args, **kwds)	Percent point function (inverse of **cdf**) at q of the given RV.
isf(self, q, *args, **kwds)	Inverse survival function (inverse of **sf**) at q of the given RV.
moment(self, n, *args, **kwds)	n-th order non-central moment of distribution.
stats(self, *args, **kwds)	Some statistics of the given RV.
entropy(self, *args, **kwds)	Differential entropy of the RV.
expect(self[, func, args, loc, lb, ub, …])	Calculate expected value of a function with respect to the distribution for discrete distribution by numerical summation.

median(self, *args, **kwds)	Median of the distribution.
mean(self, *args, **kwds)	Mean of the distribution.
std(self, *args, **kwds)	Standard deviation of the distribution.
var(self, *args, **kwds)	Variance of the distribution.
interval(self, alpha, *args, **kwds)	Confidence interval with equal areas around the median.

연속형

beta(*args, **kwds)	A beta continuous random variable.
cauchy(*args, **kwds)	A Cauchy continuous random variable.
chi2(*args, **kwds)	A chi-squared continuous random variable.
erlang(*args, **kwds)	An Erlang continuous random variable.
expon(*args, **kwds)	An exponential continuous random variable.
exponnorm(*args, **kwds)	An exponentially modified Normal continuous random variable.
exponweib(*args, **kwds)	An exponentiated Weibull continuous random variable.
exponpow(*args, **kwds)	An exponential power continuous random variable.
f(*args, **kwds)	An F continuous random variable.
gamma(*args, **kwds)	A gamma continuous random variable.
gompertz(*args, **kwds)	A Gompertz (or truncated Gumbel) continuous random variable.
invgamma(*args, **kwds)	An inverted gamma continuous random variable.
invgauss(*args, **kwds)	An inverse Gaussian continuous random variable.
invweibull(*args, **kwds)	An inverted Weibull continuous random variable.
laplace(*args, **kwds)	A Laplace continuous random variable.
logistic(*args, **kwds)	A logistic (or Sech-squared) continuous random variable.
loggamma(*args, **kwds)	A log gamma continuous random variable.
loglaplace(*args, **kwds)	A log-Laplace continuous random variable.
lognorm(*args, **kwds)	A lognormal continuous random variable.
loguniform(*args, **kwds)	A loguniform or reciprocal continuous random variable.
norm(*args, **kwds)	A normal continuous random variable.
pareto(*args, **kwds)	A Pareto continuous random variable.

rayleigh(*args, **kwds)	A Rayleigh continuous random variable.
rice(*args, **kwds)	A Rice continuous random variable.
t(*args, **kwds)	A Student's t continuous random variable.
uniform(*args, **kwds)	A uniform continuous random variable.
wald(*args, **kwds)	A Wald continuous random variable.
weibull_min(*args, **kwds)	Weibull minimum continuous random variable.
weibull_max(*args, **kwds)	Weibull maximum continuous random variable.

연속형 only 매개변인

fit(self, data, *args, **kwds)	Return MLEs for shape (if applicable), location, and scale parameters from data.
fit_loc_scale(self, data, *args)	Estimate loc and scale parameters from data using 1st and 2nd moments.

파이썬 사용예제 (이산형)

H정류장에 오는 버스 수가 평균이 시간당 10대인 포아송분포를 따른다고 하자.

오전 7시부터 오후 7시까지 H정류장에 매 10분당 오는 버스 대수에 대한 데이터를 생성해보자.

12시간*6번(시간당->매10분)=72개 데이터를 생성하는 것이다.

모수인 평균은 시간당 10대 이므로 10분당 평균은 10/6이 된다. 포아송분포의 성질에 의해

```
import scipy.stats as st
bus=st.poisson.rvs(10/6,0,72)
bus
```

데이터는 랜덤 생성되어 다를 수 있음

```
array([1, 2, 0, 2, 1, 1, 5, 2, 2, 0, 2, 0, 0, 2, 0, 2, 1, 4, 1, 3, 1, 3,
       4, 4, 1, 0, 1, 4, 0, 4, 2, 1, 4, 2, 0, 3, 2, 2, 2, 3, 0, 4, 2, 4,
       2, 3, 1, 1, 1, 1, 4, 0, 1, 2, 2, 2, 2, 1, 2, 0, 2, 0, 1, 0, 5, 1,
       2, 3, 3, 2, 1, 1])
```

데이터 기반 확률, 누적 확률 만들기

np.unique 함수에 의해 데이터의 서로 다른 관측치와 빈도를 계산하여 관측치 개별 값은 앞 array에 빈도는 뒤 array에 저장된다.

그러므로 sum(fx)를 하면 데이터 개수가 되고 fx/sum(fx)는 상대빈도, 즉 확률이 된다.

```
import numpy as np
x,fx = np.unique(bus,return_counts=True)
print (x, fx/sum(fx))
```

```
[0 1 2 3 4 5] [0.18055556 0.26388889 0.30555556 0.09722222 0.125      0.02777778]
```

$P(X=3)$ 버스가 3대 올 확률값 (이론/데이터)

```
print("P(X=3) 확률? : 데이터=%.3f 이론=%.3f" % (fx[3]/
sum(fx),st.poisson.pmf(3,10/6)))
```

```
P(X=3) 확률? : 데이터=0.097 이론=0.146
```

$P(X \leq 3)$ 버스가 3대 이하 올 누적 확률값 (이론_데이터)

```
sum(fx[0:4])/sum(fx), st.poisson.cdf(3,10/6)
```

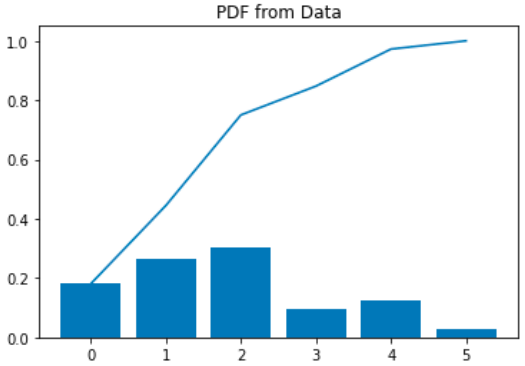
```
(0.8472222222222222, 0.9117328482652676)
```
누적분포함수

```
Fx=np.cumsum(f)/sum(f)
Fx
```

```
array([0.18055556, 0.44444444, 0.75      , 0.84722222, 0.97222222,
       1.        ])
```

확률분포함수_누적분포함수 (데이터 히스토그램) 그리기

```
import matplotlib.pyplot as plt
plt.title("PDF from Data")
plt.plot(x,FX,label='data_CDF')
plt.bar(x,fx/sum(fx),label='data_CDF')
ax.legend(loc='best', frameon=False)
plt.show()
```



모수 추정

모집단 포아송 분포의 모수는 평균이다.

```
bus.mean()
```

```
1.8055555555555556
```

파이썬 사용예제 (연속형)

H정류장에서 승객이 버스를 기다리는 평균 시간이 6분이고 지수분포를 따른다고 하자.

오전 7시를 시작으로 100명의 승객이 각각 기다리는 시간을 생성해보자.

```
import scipy.stats as st
wait=st.expon.rvs(0,6,100)
wait
```

```
array([ 0.08300885,  1.03553673,  5.42835601,  0.27120953,  7.31797018,
        9.92864099,  5.55973032, 10.02600466,  0.8591046 , 35.01591305,
        1.99091883,  4.96085829,  6.02615629,  0.06348523,  6.82640992,
        1.5091482 ,  3.92934327,  1.41385621,  5.52019968,  2.43514622,
```

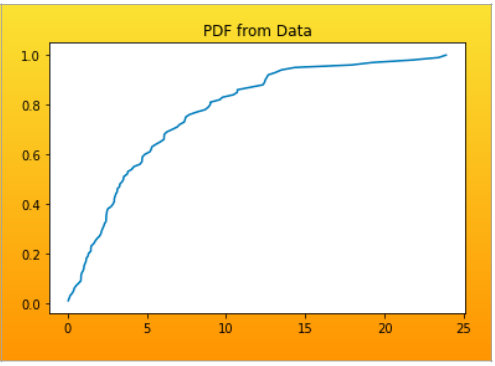
데이터 기반 누적 확률 만들기

```
import numpy as np
x,f = np.unique(wait,return_counts=True)
pdf=f/sum(f)
cdf=pdf.cumsum()
wait_f=np.column_stack((x,cdf))
wait_f[0:3,0:]
```

```
array([[0.06348523, 0.01      ],
       [0.08300885, 0.02      ],
       [0.10643829, 0.03      ]])
```

이론/실증 pdf_cdf 그리기

```
import matplotlib.pyplot as plt
plt.title("CDF from Data")
plt.plot(x,cdf,label='data_CDF')
plt.show()
```



실증데이터 모수 구하기

```
print('실증: 평균=%.3f, 표준편차=%.3f' % (wait.mean(),wait.std()))
```

```
실증: 평균=5.842, 표준편차=6.760
```

데이터

철학

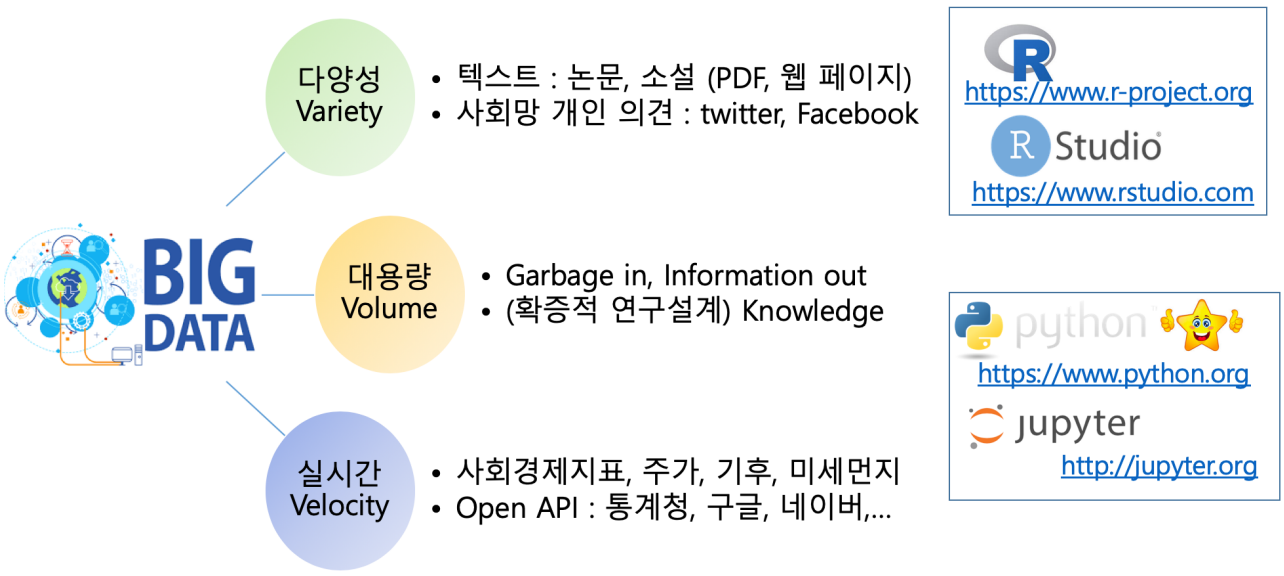
과학은 이론적 통찰 (예: 상대성 이론), 새로운 현상의 관찰 (Kepler 행성 궤도 관련 법칙)이나 경험을 (Student T-분포) 통한 새롭고 혁신적인 이론이 만들어지는 경우는 극히 드물고 대부분 관찰, 실험, 분석 등의 반복을 통해 이론이 정립된다. 벼 품종 개량, 새 의약품 개발, 화학 공정 개선 등이 실험 계획에 의한 연구 결과가 이에 해당된다.

통계 전문가는 제시된 이론을 통계적 가설이나 통계 모형으로 설정하고 관련 데이터를 수집하여 가설(모형)의 유의성을 검정하거나(confirmatory data analysis) 수집된 데이터를 탐색하여 가능한 모형이나 이론을 제시하는 역할을(exploratory data analysis) 담당하고 있다. 이처럼 (탐색적) 데이터 분석이 타 분야의 새로운 이론 발견에 기여할 수 있으려면 1)그 분야에 대한 지식 2)모형과 데이터 3)그리고 모형과 데이터의 사이클 개념을 올바로 이해해야 한다.

데이터 정의

(정의) 추론, 정보획득, 계산에 사용되는 실제의 조사, 측정된 값 (통계) (웹스터), 정보를 가진 숫자의 모임

통계학에서 분석 대상인 데이터는 행과 열로 이루어져 있으며, 행은 개체 subject 관측치, 열은 관심 특성, 변수 variable로 이루어진 숫자 행렬이다.행의 첨자는 개체를 나타내고, 열의 첨자는 변수를 나타낸다.



데이터 종류

고전적(분석적) 정의

통계학에서는 변수 variable, 사회과학방법론에서는 scale라 한다.

질적 Qualitative, Non-metric (분류형 변수, classified, 범주형 categorical) : 개체를 분류하기 위해 측정된 변수를 의미 하며 성별, 결혼여부 등이 그 예이다.

- 명목 (nominal): 개체를 구분하거나 분류할 목적 (예) 성별, 결혼여부, 직업, 거주시도
- 순서, 서열 (ordinal): 개체를 일정한 순위순서를 가진 분류, ? 성적(A, B, ..) 소득수준(상, 중, 하),

양적 Quantative, Metric (측정형 변수, measurable) : 실험 개체의 측 정 가능한 특성을 측정한 변수로 키, 몸무게, 평점, IQ, 교통량, 사망자 수가 그 예이다. 연속형 변수는 모두 측 정형 변수이고 이산형 변수 중 측정형 변수가 있을 수 있다. 예) 교통량

- 구간, 등간 interval : 0 의미 없고 배율도 의미 없음 (예) 온도, 리커트 척도
- 비율척도 ratio : 0이라는 숫자의 의미가 있고 배율의 의미 존재 (예) 소득, 수능성적, 키, …

	구간	비율	순서	명목
빈도표	X	X	X	X
순서 있음		X	X	X
최빈값	X	X	X	X
평균	X	X		
중위수	X	X	X	
+, - 가능	X	X		
곱셈, 나누셈		X		
0의 개념, 배율		X		

시간적 정의

자료가 시간적 순서를 가지면 이를 시계열 자료(time series)라 하고 그렇지 않은 경우를 횡단면 자료(Cross-section: 일정 시간에 한꺼번에 조사)라 한다.

- 횡단 변수 cross section : 일정 시점의 조사 데이터
- 종단변수 time series : 시간적 순서를 갖는 데이터 - 대신 로 표현, 경제 지표(환율, 수출량)나 기업의 연차별 자료(연도별 매 출액), 연도별/월별 청년 실업율

포멧

- 숫자 : 명목형 데이터는 문자(범주)이지만 분석 시 class 변수로 설정하여 숫자처럼 사용한다.

비정형 No-Structural

- 문자 : 텍스트 마이닝, 자연어 처리, word cloud
- 음성 : 오디오 파일 포멧을 데이터로 변환/역변환 가능 (convert audio format into data format)

- 이미지 : 이미지 파일 (x, y) 좌표나 6자리 숫자 단위로 변환 가능

dataframe, series, matrix, array

Series : 데이터 형태 (integers, strings, floating point numbers, Python objects, etc.) 관계없이 레이블을 가진 하나의 열이다.

- 행인덱스를 가지고 있다.
- numpy 사용되는 매개 argument 사용가능하다.

Dataframe : 이차원 레이블 데이터 구조 (열에는 서로 다른 데이터 형태를 가질 수 있음) 엑셀 시트나 SQL 데이블과 동일하다.

Matrix : 데이터만 저장되어 있으며 사칙연산이 가능한 측면에서 배열 array와 동일하나 2차원이다.

Array : 데이터만 저장되어 있으며 사칙연산이 가능하다. 반드시 numpy.array()로 선언해야 한다. N차원 데이터구조이다.

- list : 상이한 형태의 데이터 저장 가능, series와 유사하나 행 인덱스 없음
- ndarray는 numpy에서 지원하는 표준형인 벡터/행렬/텐서 를 저장한다.
- 많은 numpy 함수가 matrix가 아니라 ndarray를 반환한다.
- 요소 간 연산과 선형대수 연산에 대해선 명확히 구분되어 있다.
- 표준 벡터나 열벡터/행벡터를 표현할 수 있다.

상관관계 correlation 인과관계 casual relationship

상관관계

변수들간의 직선 관계가 있는지에 대한 척도로 상관계수를 이용한다. 산점도를 활용하여 두 확률변수의 함수관계를 예상할 수 있다.

인과관계

인과 관계 모형은 통계분석에 의해 검증되는 것이지 발견하는 것은 아니다. 모형 설정은 이론, 경험적 타당성에 근거하여 이루어진다.

- X - 독립변수, 요인(처리효과), 예측변수, 설명변수, 내생변수 : 원인이 되는 변수
- Y - 종속변수, 반응변수, 목표변수, 외생변수 : 영향을 받는 변수

인과 관계는 이론적, 경험적 타당성에 근거하여 데이터 **수집 전에 이론이나 경험에 의해 설정**되는 것이지 분석 후 설정되는 것은 아니다.예를 들어 보자. 대학생 40명의 GPA성적과 용돈을 조사하였다. GPA성적-Y, 용돈을 X로 하여 회귀분석을 실시한 결과 유의한(significant)가 있다고 해서 용돈이 수능성적에 영향을 미친다고 할 수 있나? No, 인과관계 설명은 또 다른 이야기임, 최종 인과관계는 사회실험 조사 필요

고전적인 예제 - 흡연과 암의 인과관계 : 횡단시점 분석(코호트 분석), 암 환자의 흡연비율과 일반인 흡연비율 차이 비교 - 그러나 숨은 변인 효과

association (연관)이 인과 (causation)를 의미하는 것은 아님 - 두 변수의 관계를 분석할 때 고려하지 않았으나 관계에 영향을 미치는 변수를 lurking(숨은, 잠재) 변수로 인한 것임, 흡연과 암의 관계에서 소득수준이 잠재변인일 수 있음, 소득수준 효과를 제외하면 흡연과 암의 (인과)관계가 존재하지 않을 수 있음

데이터와 확률변수 random variable

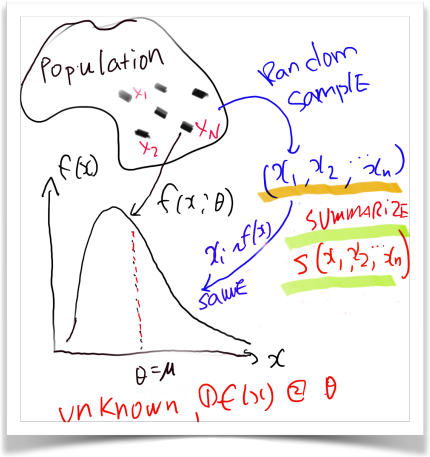
통계분석의 대상이 되는 데이터는 변수들로 구성되어 있는데, 변수의 관측결과와 숫자를 일대일 매칭한 함수를 확률변수라 한다. 변수의 관측치가 숫자인 경우는 데이터 값 자체가 확률변수이다.

확률변수가 가질 수 있는 값과 대응하는 확률을 일대일 매칭한 함수를 확률분포함수(pdf probability density function)라 한다. 데이터의 히스토그램이 확률분포함수와 동일하다.

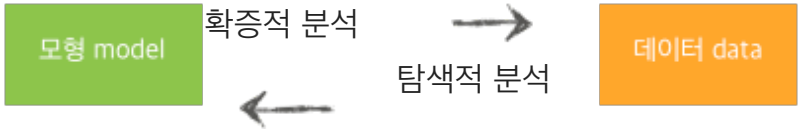
통계적 방법은 데이터에 포함된 정보 모두를 가져오는 것이 아니라

데이터 요약 개념

- 모집단으로부터($X \sim f(x;\theta)$) 얻은 확률표본(데이터)은 $(x_1, x_2, \ldots, x_n)$는 모집단의 분포와 동일하다.
- 모집단에서 궁금한 것은 확률분포함수 $f(x)$와 모수(θ)이다.
- 모수에 대한 정보는 확률표본, 데이터의로부터 계산될 수 있는데, f(x)는 그래프 graphical 요약으로부터 모수에 대한 정보는 숫자 numerical 요약으로 얻는다.
- 통계적 방법론에서 데이터 종류는 "숫자형, 정량변수", "분류, 범주형, 정성변수", 2개로 나뉘고, 변수의 종류에 따라 요약방법이 정해진다.



모형과 데이터 사이클



과학에서 이론이 제안되고 데이터 분석이 이루어지는 경우보다는 데이터로부터 새로운 이론이나 모형을 도출하는 경우가 많고 탐색적 자료 분석에 의해 제안된 이론이나 모형은 다시 confirmatory 방법에 의해 유의성이(significance) 검증되므로 모형과 데이터는 순환 사이클을 갖는다.

통계적 모형은 과학적 진실이기 보다는 분석 대상이 되는 사실(현황)의 대표적 모형이다. 예를 들어, 회귀모형에서는 ($Y = \alpha + \beta X + e$) 선형함수(모형)이 설명하지 못하는 오차항(e)이존재하고 이 오차항은 평균 0, 분산 σ^2 인 정규분포를 따른다고 가정한다.

연역적 방법 (deductive reasoning)

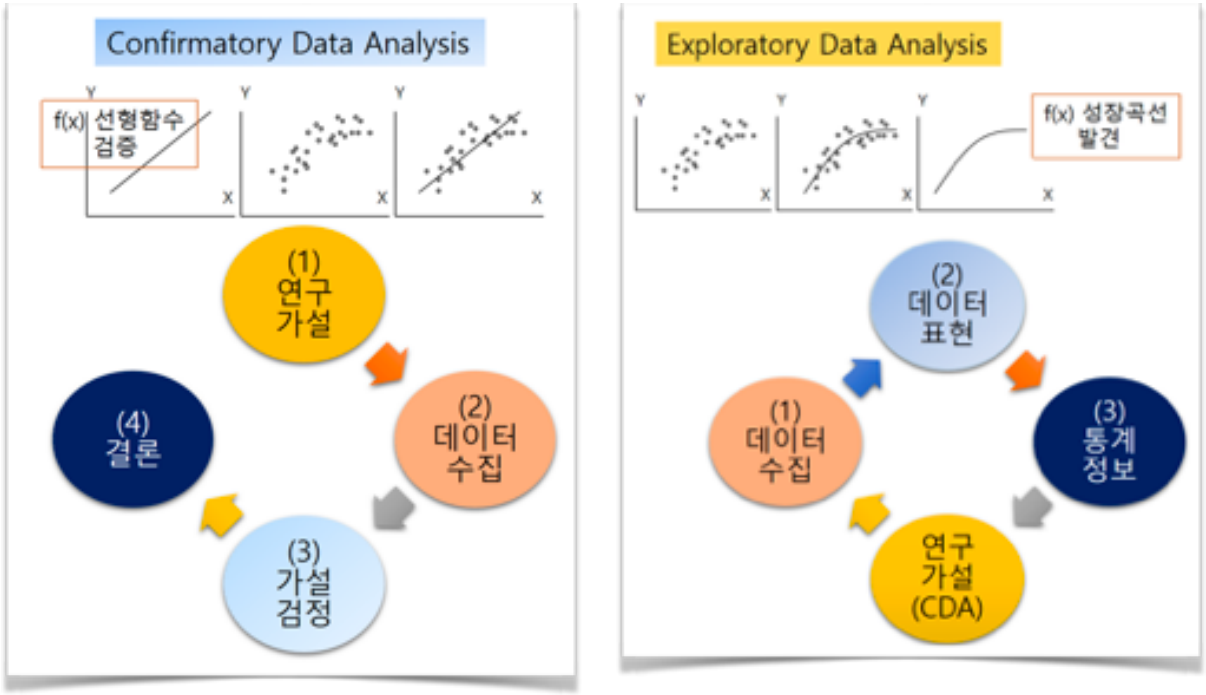
Confirmatory(확증적) Data Analysis 과학철학자 Popper(11955)는 "이론은 직관에 의해서만 얻어질 수 있다"고 주장해 연역적 방법의 타당성을 강조하였다.

연구가설 설정 -> 데이터 수집 -> 가설 검증 및 결론

귀납적 방법 (inductive reasoning)

1977년 John W. Tukey 제안 탐색적 데이터 분석(EDA: Exploratory Data Analysis) 방법

(1)수집된 데이터가 가진 정보를 숫자 요약과 그래프를 이용하여 찾아내거나 (2)데이터를 보다 유용하게 만들기 위하여 데이터를 재표현(re-expression) 하여 정보 획득 => Data Mining => Big Data



데이터 확률분포

코로나 한국 확진자 데이터 확률분포

데이터 불러오기

원 데이터 불러와서 원데이터는 dateRep 는 오브젝트로 되어 있어 시간 포멧으로 변형하였음

```
import pandas as pd
df=pd.read_csv('https://opendata.ecdc.europa.eu/covid19/casedistribution/csv')
#df['dateRep'] =pd.to_datetime(df.dateRep)
```

시간에 의해 데이터 정렬

```
df_kor=df[df['countryterritoryCode']=='KOR']
df_kor.sort_values(by=['year','month','day'],inplace=True)
df_kor.head(3)
```

누적 확진자수, 사망자수 데이터 만들기

```
df_kor['cum_cases']=df_kor['cases'].cumsum()
df_kor['cum_deaths']=df_kor['deaths'].cumsum()
```

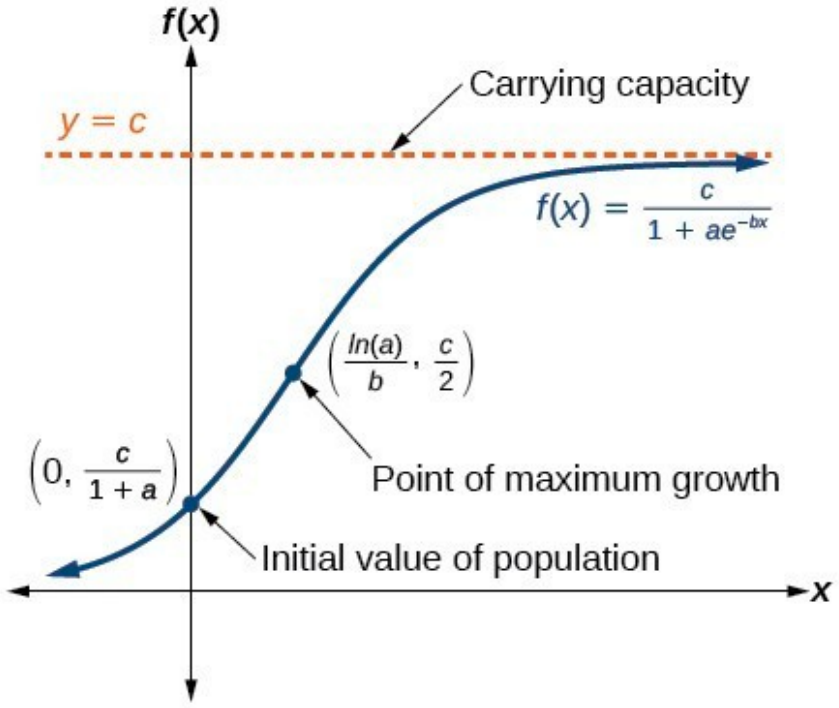
원데이터 행 인덱스 번호를 0부터 시작하게 바꾸고 원 인덱스는 삭제하였음

```
df_kor.reset_index(inplace=True)
df_kor.drop(['index'],axis=1,inplace=True)
```

```
from plotly.subplots import make_subplots
import plotly.graph_objects as go
from matplotlib import pyplot as plt
import plotly.express as px
fig=make_subplots(rows=2,cols=2,subplot_titles=("cases","cum_cases","deaths","cum_deaths"))
fig.add_trace(go.Scatter(x=df_kor['dateRep'], y=df_kor['cases']),row=1, col=1)
fig.add_trace(go.Scatter(x=df_kor['dateRep'], y=df_kor['cum_cases']),row=1, col=2)
fig.add_trace(go.Scatter(x=df_kor['dateRep'], y=df_kor['deaths']),row=2, col=1)
fig.add_trace(go.Scatter(x=df_kor['dateRep'], y=df_kor['cum_deaths']),row=2, col=2)
fig.update_layout(height=600, width=1200,
title_text=df_kor['countriesAndTerritories'].iloc[1])
fig.show()
```

【실습문제】
누적 확진자의 분포를 '지수분포', '정규분포'에 접근하여 보자.

Logistic growth model



$$f(x) = \frac{c}{1 + ae^{-bt}}$$

$f(x)$: 시점 x에서의 확진자 수

$c/(1+a)$: 확진자 초기값

c : 확진자 (예상) 최대값

b : 성장율 growth rate 에 의해 결정되는 상수

growth rate 추정방법

지수성장모형 : $N_t = n_0 e^{bt}$ 활용하여 추정한다.