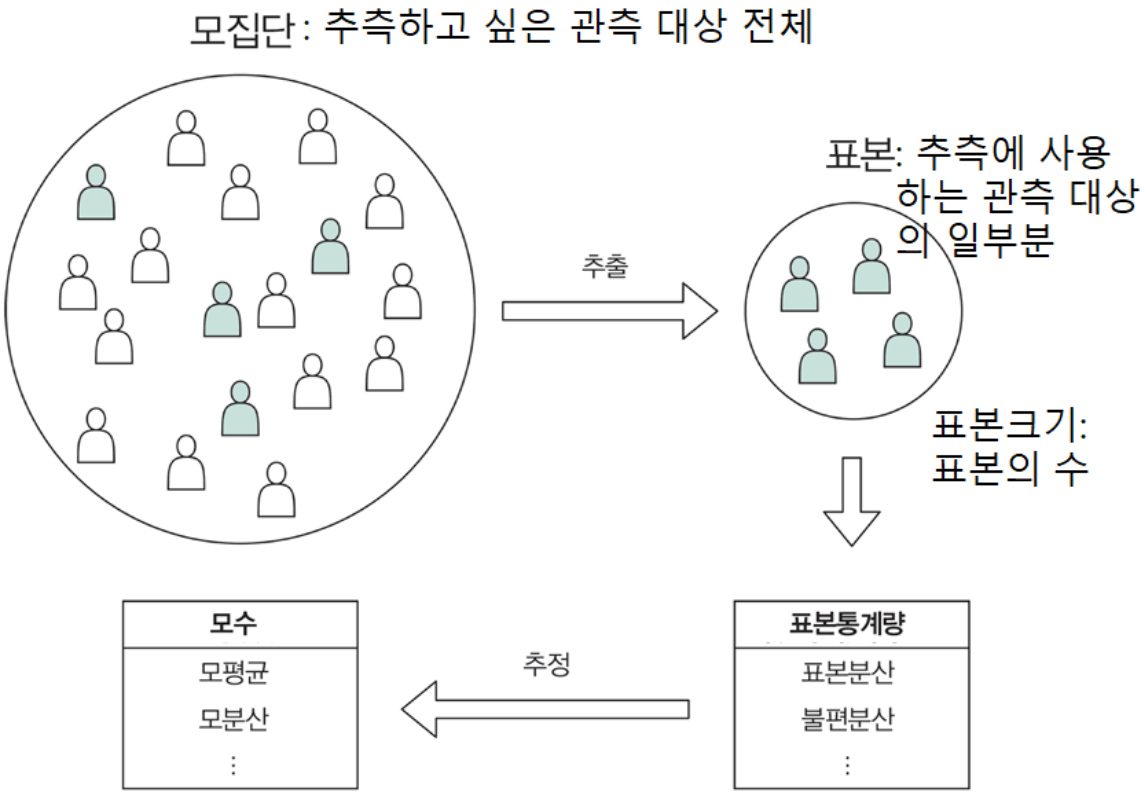
확률과 통계

도입

모집단과 표본

모집단



12

1.1 곱의 원리

- 원소의 개수가 $n_1, n_2, \cdots, n_k$ 인 집합 $A_1, A_2, \cdots, A_k$ 에서 각각 한 개의 원소를 택하여 나열한 순서열의 개수는 $n_1 n_2 \cdots n_k$이다
- 예) 서울에서 대전으로 가는 방법은 세가지 방법이 있고 대전에서 부산으로 가는 방법은 두가지가 있다. 그러면 서울에서 대전을 거쳐 부산으로 가는 방법의 수는 몇가지 인가?

순열

n개의 원소를 가진 집합에서 k개의 서로 다른 원소를 택하여 이룬 순서열 $(z_1, z_2, \cdots, z_k)$을 n 개에서 k개를 택한 순열 $({}_nP_k)$이라 하고, 중복올 허락 하여 이룬 순서열올 중복순열 $({}_n\Pi_k)$이라 한다

- ${}_nP_k = n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}$
- ${}_n\Pi_k = n^k$
- 참고 $n! = n \cdot (n-1) \cdots 2 \cdot 1, 0! = 1$

- 예) 문자 a,b,c,d,e,f 중에서 세 개를 선택하여 만들 수 있는 단어의 수는 몇 가지 인가?
- 예) 10명 중에서 각각 다른 사람으로 회장, 부회장, 총무를 선출할 수 있는 방법은 몇 가지 인가?

중복순열

n_1개 개체들이 같고, n_2개 개체들이 같고, $\cdots$, n_r 개 개체들이 같은 총 n개의 순열의 수는 다음과 같다.

- $\frac{n!}{n_1!n_2!\cdots,n_r!}$ (단, $\sum_{i=1}^{r} n_i = n$)
- 예) 문자 banana로 만들 수 있는 단어의 수는 몇가지 인가?

복원추출과 비복원 추출

- 복원추출 : 원래 상태에서 r개 선택
  - $nn\cdots n = n^r$
- 비복원 추출 : 하나씩 차례로 r 개 선택
  - $n(n-1)\cdots(n-r+1) =_n P_r = \frac{n!}{(n-r)!}$

- 예) 52장의 카드 한패에서 3장을 선택하는 각각의 경우의 수를 구하여라
  - 복원 추출의 방법
  - 비복원 추출의 방법
  - 동시 추출의 방법

조합(Combination)

n개의 원소를 가진 집합에서 k개의 서로다른 원소를 택하여 이룬 집합 $\{z_1, z_2, \cdots, z_k\}$를 n개에서 k개를 택한 조합 ($_nC_k$)이라 한다. 즉, 선택된 k개의 개체는 순서에 무관하다.

- $_nC_r = \frac{_nP_r}{r!} = \frac{n!}{r!(n-r)!}$

- 예) 4개의 문자 a,b,c,d 에서 3개를 택하는 조합을 구하여라

표본

표본공간

확률을 계산하려연 확률실험을 행한다. 이때, 모든 가능한 실험결과들의 집합을 표본공간이라 하고 관심 있는 실험결과들의 집합을 **사상** 또는 **사건**(event) 이라고 한다 예를 들연 주사위를 1회 던졌올 때 표본공간 $S = \{1, 2, 3, 4, 5, 6\}$ 이고 사상 $A = \{3\}$ 으로 항상 사상은 표본공간의 부분집합이다

- **표본공간** 표본추출이나 통계실험올 통해 얻어진 가능한 모든 결과로 S, Ω 로 표시한다.
- **사 상** : 표본공간 S의 부분집합을 사상 또는 사건이라 하며 A, B, C 등 을 이용해서 표시한다.
  - 사상의 형태에 따른 사건의 분류
    - $A \cup B$ 는 A가 발생하거나 B가 발생(또는 양쪽 모두 발생)한 사상이다.
    - $A \cap B$는 A와 B가 동시에 발생하는 사상이다.
    - A^c는 A 가 발생하지 않는 사상이다.
    - $A \cap B = \emptyset$일때 서로 배반적 이라고 부른다. (동시에 일어나지 않음)

표본추출(Sampling)

- 가정에 따른 분류
  - 확률적 표본추출방법
    - 동일한 확률을 가정하고 표본을 구성
  - 비확률적 표본추출 방법
    - 확률과는 상관없이 자원 또는 무작위로 추출

- 추출 방식에 따른 분류
  - 무작위 추출 (임의 추출) : 임의로 표본을 추출하는 방법
  - 복원추출 : 여러 차례 동일한 표본을 선택하는 방법
  - 비복원 추출 : 동일한 표본은 한 번만 선택하는 방법

확률 공리와 성질

- 확률의 고전적 정의
  - 표본공간 S가 유한이고 각가의 원소가 일어날 가능성이 같을때 사상 A의 확률로 정의되며 $|A|$는 사상 A의 원소의 개수를 나타 낸다.
    $P(A) = \frac{|A|}{|S|}$
- 확률의 공리적 정의
  - $(P_1)\ 0 \leq P(A) \leq 1$
  - $(P_2)\ P(S) = P(\Omega) = 1$
  - $(P_3)\ A_1$와 A_2가 서로 배반인 사상에 대하여 $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ 를 만족할때 P를 **확률함수(probability function)**라 하며 $P(A)$를 사상 A에 대한 **확률(Probaility)** 이라고 한다.

조건부 확률

조건부 확률은 한 사상이 일어났다는 조건 하에 다른 사상이 일어날 확률이다. 사상 A가 발생하고 난 후 사상 B가 발생한 확률, 즉 표본 공간 S에서 축소된 공간 A에 관한 B의 상대 확률로서 $P(B|A)$로 표기한다.

- 조건부 확률의 정의
  - $P(B|A) = \frac{P(A \cap B)}{P(A)}$ (단, $P(A) > 0$),
    혹은 $P(A \cap B) = P(A)P(B|A)$
- 예) 주사위를 한번 던질 때 는 B짝수, A는 눈이 4 이상의 사상이라 하자 이때, 확률 $P(B|A)$를 구하여라
- 예) 어느 대학에서 40%가 여성이고, 그중 10%가 O형의 혈액형을 가졌다. 그 대학에서 랜덤하게(무작위) 한 사람을 뽑았을때 O형을 가진 여성일 확률을 구하여라

조건부 확률의 일반화

- $P(A_1) > 0$와 $P(A_1 \cap A_2) > 0$인 사상 A_1, A_2 그리고 A_3에 대하여 다음이 성립한다.
  - $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)$
  - 일반적으로 $P(A_1 \cap A_2 \cap A_3 \cdots \cap A_k) > 0 \ (1 \leq k \leq n-1)$ 인 사상 $A_1, A_2, \cdots, A_n$ 에 대하여
    $P(A_1 \cap A_2 \cap A_3 \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$

- 예) 52장의 카드에서 한 장씩 뽑을때 순서대로 2, 3, 8, 8이 나올 확률을 구하여라.

전체 확률 법칙(The Total Probaility rule)

사상 $A_1, A_2, \cdots, A_n$이 표본공간의 분할이고 $P(A_i) > 0$이면, 임의의 사상 B에 대하여 다음이 성립한다.

$P(B) = P(A_1 \cap B) + P(A_2 \cap B) \cdots P(A_n \cap B)$
$= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n)$

- 두 개의 상자가 있다. 첫 번째의 상자안에는 빨간 공 2개, 흰 공 5개 가 들어있다. 두 번째 상자안에는 빨간 공 3개, 흰 공 4개가 들어있다. 첫 번째 상자에서 하나의 공을 꺼내어 두 번 째 상자에 넣고 두 번째 상자에서 하나의 공을 꺼낼 때 빨간 공이 나올 확률을 구하여라

베이즈 정리

- 조건부확률을 구하는 공식을 베이즈 정리(Bayesian rule) 라고 한다.
  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

나이브 베이즈 분류

$A_1, A_2, \cdots, A_n$이 S의 분할이고 B를 임의의 사건이면

이때, 사건A_i 가 서로 배타적이고 완전하다고 하자.

- 서로 배타적(교집합이 없다) $Ai \cap Aj = \emptyset$
- 완전(합집합이 표본공간)$A1 \cup A2 \cup \cdots = \Omega$
  $P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)+\cdots+P(A_n)P(B|A_n)}$

(단, $P(A_i) > 0, P(B) > 0$)

- 특정 사건 A_1에 대한 조건부 확률

$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B)} = \frac{P(B|A_1)P(A_1)}{\sum_i P(A_i \cap B)} = \frac{P(B|A_1)P(A_1)}{\sum_i P(B|A_i)P(A_i)}$

- 베이지안 필터
- 확률을 이용해 문제를 분류해야 하는경우 사용

- 예
  - URL 링크 유무에 따른 스팸 메일 판단
  - 혈액 수치에 따른 질병 유무 확률
  - 문서 분류 (스포츠, 정치, 연예 등)

- 예) 한 질병에 대한 혈액검사의 적중률은 95%이다. 즉, 실제 질병이 있는 사람의 혈액검사 결과가 양성으로 나타날 확률이 0.95 이다. 또한, 실제로는 질병이 없는 사람의 혈액검사 결과가 양성으로 나타날 확률이 0.01% 이다. 이제, 이 질병의 감염률이 1000명 중에서 5명 꼴로 회귀할 때, 혈액검사의 결과가 양성으로 나타난 사람이 실제로 이 질병에 걸려 있을 확률을 구하여라.

사건의 독립

두 사상 A, B에 대하여 사상 B는 사상 A가 발생하거나 또는 발생하지 않거나 영향을 미치지 않는다. 이때 **사상 B는 사상 A 와 독립** 이라 한다.

다시 말하면 사상 B의 확률은 사상 A에 대한 B의 조건부 확률과 같다. 즉, $P(A|B) = P(A)$이거나 $P(B|A) = P(B)$이다.

- 종속
  - $P(A \cap B) = P(A) \cdot P(B)$이면 사상 A와 B는 독립이다. 만일 이 식이 성립하지 않으면 **종속**이라 한다.

- 동전을 던지는 사건
- 예) A가 목표물을 명중할 확률은 $\frac{1}{2}$이고, B가 명중할 확률은 $\frac{2}{3}$이다. 두 사람이 목표물에 사격할 때 명중할 확률은 얼마 인가?

- 독립 시행을 반복하여 나타낸 결과를 분포로 나타내면 어떻게 될까?

이항 확률(Binomial Probability)

사상 A가 n회 독립시행으로 매번 확률 p와 동일하다고 가정하자. $K_A(n)$는 n회 독립시행에서 사상 A가 발생 할 횟수일 때 다음의 식이 성립하면 **이항 확률분포**라 한다.

- $P(K_A(n)=m)=\binom{n}{m}p^m(1-p)^{n-m}(0\leq m\leq n)\backslash$
- Notation : $B(n,p)$

- 예) 주사위를 6회 독립적으로 던질 때 눈 6이 나오는 횟수의 확률을 구하여라

확률 변수

확률 실험에서 관심이 되는 표본공간의 원소보다는 그에 관련된 수치적 함수인 경우가 많이 있다.
한 쌍의 주사위를 던지는 시행에서 표본공간은 $S=\{(i,j)|i,j=1,2,3,4,5,6\}$ 이고 관심의 대상은 두 주사위의 눈의 합이나 차, 즉 $X=i+j,|i-j|$ 가 된다. 이와 같은 확률실험의 수치부여를 **확률변수**라 부른다.

- 예) 두 개의 주사위를 던지는 실험에서 확률변수 X는 두 눈의 합이라고 하자, 그러면 확률변수 $X=\{2,\cdots,12\}$ 이고 각각의 확률을 구하여라

이산 확률 변수

X가 유한 개 혹은 가산 무한개 값을 취할 수 있는 확률 변수인 경우 **이산확률변수**라 하고 $X=\{x_1,x_2,\cdots\}$로 나타내고 x_i에서의 확률 $f(x_i)=P(X=x_i)$를 X의 **확률질량함수** 라 부르고 확률 분포표는 다음과 같다.

X	x_1	x_2	x_3	$\cdots$	합
$f(x_i)=P(X=x_i)$	$f(x_1)$	$f(x_2)$	$f(x_3)$	$\cdots$	1

- 확률 질량함수 f의 특성은 다음과 같다.
  - $0\leq f(x_i)\leq 1$
  - $\sum_{i=1}^{\infty}f(x_i)=1$

- $P(a \leq X \leq b) = \sum_{a \leq x_i \leq b} f(x_i)$

누적 분포함수

이산확률 변수의 누적 분포 함수는 다음과 같이 정의 한다.

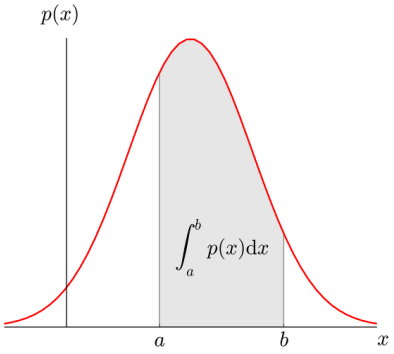
- $F(x) = P(X \leq x) = \sum_{x_i \leq x} P(x_i)$일 때 F는 **누적분포함수** 라 한다.
- 분포한수 F는 다음 성질을 만족 한다.
  - $F(-\infty) = 0, F(\infty) = 1, 0 \leq F(x) \leq 1$
  - $P(a < X \leq b) = F(b) - F(a)$

- 예) 두 개의 동전을 던지는 시행에서 앞면이 나오는 횟수를 X라 할 때, 누적분포함수를 구하여라.

연속확률변수 (Continuous random variable)

X가 실수 전체 혹은 (수직선 상의)어느 구간에 포함되는 실수 전체의 값을 취하는 연속성을 가지는 확률 변수인 경우 **연속확률변수**라고 한다.

- $P(a \leq X \leq b) = \int_a^b f(x)dx$ 는 a 에서 b 사이의 사건이 발생될 확률을 의미 한다.



- 여기서 f는 X의 **확률밀도함수** 라 하고 다음 조건을 만족한다.
  - $0 \leq f(x)$

- $\int_{-\infty}^{\infty} f(x)dx = 1$

확률 분포

- 이산 확률 분포
- 연속 확률 분포

이산 확률분포

확률변수 X가 유한 개 또는 가산 무한개의 값 $x_1, x_2, \cdots$ 만을 취한수 있을때, 이 확률변수 X를 이산 확률변수라 하고 X의 분포를 **이산분포** 라 한다.

따라서 $P(x_i) = P(X = x_i) = f(x_i)$ 을 **확률질량함수** 라 부르며 모든 x_i에 대하여

$0 \leq P(x_i), \sum_{i=1}^{\infty} P(x_i) = 1, F(a) = \sum_{x_i \leq a} P(x_i)$

이 성립한다

- 예) $P(1) = \frac{1}{2}, P(2) = \frac{1}{3}, P(3) = \frac{1}{6}$ 일 때 누적 분포 함수를 구하여라

이산확률 분포 - 베르누이 분포

- 베르누이 실행
  - 서로 반대되는 사건이 일어나는 실험을 반복적으로 실행하는것
  - 아들 아니면 딸, 홀수 아니면 짝수 등
- 베르누이 분포
  - 베르누이 시행을 확률 분포로 나타낸것
  - 성공확률이 p 라면 실패확률은 $1 - p$ 이다.
    - 두개의 결과가 서로 배타적이고 성공할 확률이 시행때 마다 똑같다.
  - 베르누이 분포에서의 기댓값과 분산
- 베르누이 분포에서의 기댓값과 분산
  - $\mu = E(X) = p$

- $\sigma^2 = \operatorname{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right] = E(X^2) - E(X)^2$=$p - p^2 = p(1 - p)$

이산확률 분포 - 이항 분포

이항 확률변수

- 확률질량함수가 다음과 같을 때 확률변수 X를 **이항 확률변수** 라 한다.
  $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} (x = 0, 1, \cdots, n)$
- n이 충분히 크고 P가 충분히 작은 경우
- 예) 공정한 동전을 4회 던졌을 때, 2번의 앞면과 2번의 뒷면이 나올 확률을 구하여라.
- 예) 구매를 자주했을때 마음에 드는 물건을 구매할 확률은 10%이다. 약 49개의 제품을 구매했을때 2개의 제품이 마음에 들 확률을 구하여라

이산확률 분포 - 푸아송 분포

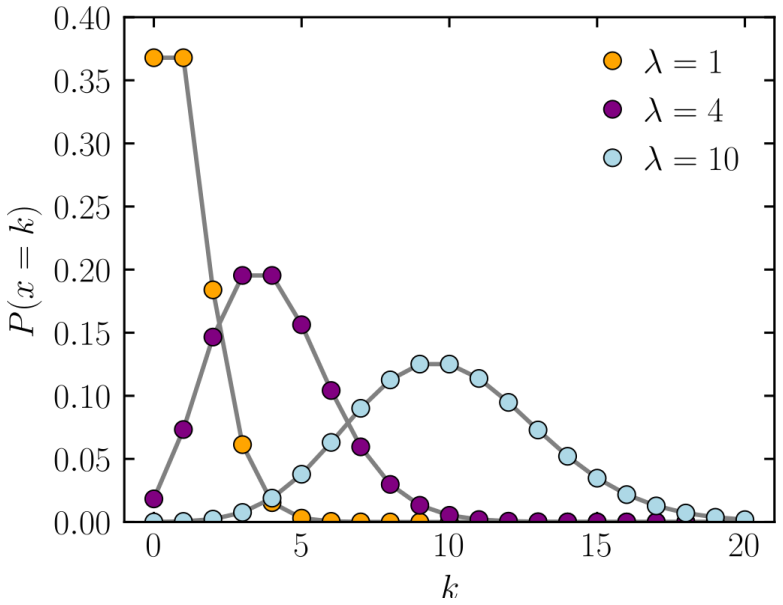
정해진 시간 안에 어떤 사건이 일어날 횟수에 대한 기댓값을 λ 라고 했을 때, 그 사건이 x 회 일어날 확률은 다음과 같다.

$$P(X = x) = f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- 일반적으로, n\geq 20n
  ≥20이고 p\leq 0.05p
  ≤0.05이면 어느 정도 충분하고, n\geq 100n
  ≥100이고 np\leq 10np
  ≤10이면 매우 훌륭하다고 여겨진다.

- 특정사건이 발생할 가능성이 매우 드문 경우의 확률분포
  - 화장실 변기에 에어팟을 떨굴 확률
  - 기마대의 기병이 낙마사고를 당할 확률
  - 지하철을 탈때 바로 도착할 확률
  - 일 주일 동안 오는 보험회사에 접수되는 사망 보험금 청구건수
  - 공장에서 불량품이 생성될 확률

- λ에 따른 확률질량 함수 값



- 예) 세 쌍둥이가 태어날 확률이 0.0001이라 할 때 10000명의 신생아 중에서 적어도 4쌍 이상이 세 쌍둥이가 될 확률을 구하여라
- 예) 공장에서 생산된 물품의 2%가 불량품 이라 하면 100개의 물품 중 3개가 불량품일 확률을 구하여라

연속확률 분포

- 시간, 높이, 무게와 같은 측저이는 어떤 구간에 있는 연속적인 값이다. 이런값들을 연속성의 구간 위에서 모든 점을 가질수 있다. 그 특성은 다음과 같다.

(i) $P[X\in(-\infty,\infty)]=\int_{-\infty}^{\infty} f(x)\,dx=1$ (즉, 면적=1)

(ii) $P(a\le X\le b)=\int_a^b f(x)\,dx$

(iii) $f(x)\ge 0$

(iv) $P(a\le X\le b)=P(a<X\le b)=P(a<X<b)=P(a\le X<b)$

$(\because\ P(X=a)=\int_a^a f(x)dx=0)$

(v) $F(a)=P[X\in(-\infty,a]]=\int_{-\infty}^{a} f(x)dx$

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

연속확률 분포 - 균등분포

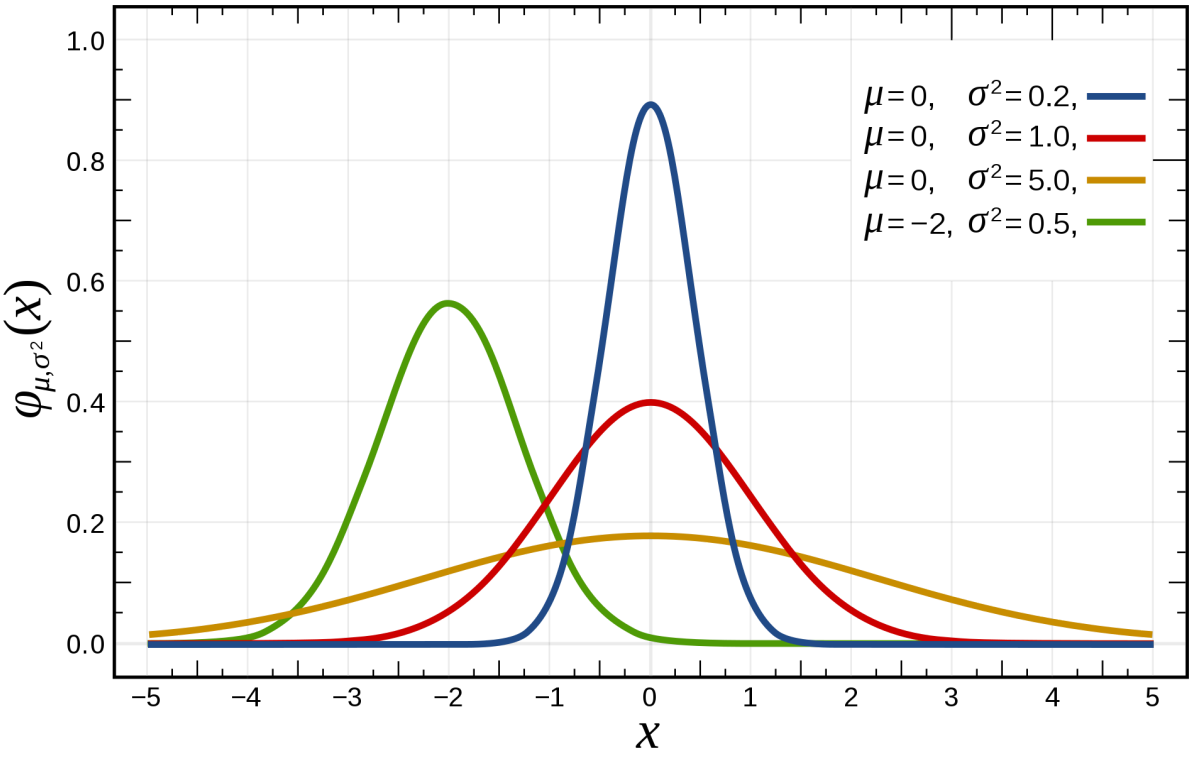
- 확률밀도함수가 다음과 같을 때 확률변수 X 를 구간 (a,b)에서 **균등 분포** 라고 한다.

$$f(x)=P(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b\\ 0, & \text{그 이외}\end{cases}$$

[Note] (i) $f(x)\geq 0$

(ii) $\int_{-\infty}^{\infty} f(x)\,dx=\int_{a}^{b}\dfrac{1}{b-a}\,dx=1$

연속확률 분포 - 정규분포 (normal distribution)



- **가우스 분포**Gaussian distribution)는 연속 확률 분포의 하나이다.
- 표본분포중 가장 단순한 형태

- 어떤 사건이 일어나는 빈도(frquency)를 계산하여 그래프로 나타내면 평균을 중심으로 좌우가 대칭되는 분포
- 정규분포의 확률밀도 함수는 분포의 평균과 분산에 영향을 받는다.

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

- 예) 어느 방송국의 연속극 시청률이 20%였다. 프로그램을 새 편성 한 후에 100명의 시청자를 임의로 뽑아 이 연속극의 시청여부를 질문하였다
  - 이 연속극의 시청률이 전과 동일하다면 100명 중 15명 이하가 시청할 확률을 구하여라
  - 시청자가 25명 보다 많을 확률을 구하여라.

- **표준 정규분포**
  - A는 수학 80점, B는 영어 90점 누가 공부를 더 잘하는가?
    - 무엇을 기준으로 비교 해야할까?
  - 정규분포의 표준화가 이루어진 상태
  - 평균 = 0, 분산 = 1
  - 서로 다른 단위의 자료들을 비교하기 위해서는 표준화가 필요하다.
  - 표준화
    - X(원점수) 를 Z-Score 로 정규화 함으로써 평균이 0, 분산이 1인 표준 정규분포를 얻을수 있다.
    - $Z = \frac{X - \mu}{\sigma}$
- 추정시에는 30개 이상의 표본일때 사용함