통계 개요

통계란 분석하고자 하는 집단에 대해서 조사하거나 실험을 통해서 얻는 자료 또는 이의 요약된 형태를 말함 통계학이란 불확실한 상황에서 효과적인 의사결정을 할 수 있도록 수치자료를 수집, 정리, 표현하고 분석하는 이런과 방법을 연구하는 학문

통계분석이란 특정집단을 대상으로 자료를 수집하여 대상집단에 대한 정보를 구하고, 적절한 통계분석 방법을 이용 하여 의사결정(통계적 추론)을 하는 과정을 말함

표본

단순 랜덤 추출법

계통 추출법

군집 추출법

층화 추출법

확률과 확률분포

확률

발생 가능한 모든 사건들의 집합 표본공간에서 표본공간의 부분집합인 특정 사건 A가 발생할 수 있는 비율을 나타내는 값으로, 0과 1 사이의 값이며, 가능한 모든 사건의 확률의 합은 항상 1이다.

P(A) =특정 사건 A의 경우

전체 사건의 경우 (표본공간)

조건부확률

특정 사건 A가 발생했다는 것이 사실이라는 전제하에 또 다른 사건 B가 발생할 확률을 나타낸 값으로, 0과 1 사이의 값을 갖는다.

P(B|A) =P(BnA)

P(A)

독립사건과 배반사건

- 독립사건

- 주사위를 2번 던지는 시행에서 첫 번째로 나오는 눈의 수가 두 번째 주사위에 영향을 주지 않는 것처럼 서로에게 영 향을 주지 않는 두 개의 사건을 독립이라고 함

- 조건부 확률에서 두 사건 A와 B가 독립인 경우에는 A가 발생했을 때를 가정하더라도 B의 확률은 변하지 않기 때문 에 다음의 식이 성립

PBA = PA

따라서 두 사건 A와 B가 독립이면 아래 식이 성립

P(AnB) = P(A)P(B)

- 배반사건

- 두 사건 A와 B에 대하여 교집합, 즉 공통된 부분이 없는 경우를 배반사건이라고 한다.

- 동시에 일어날 수 없는 사건

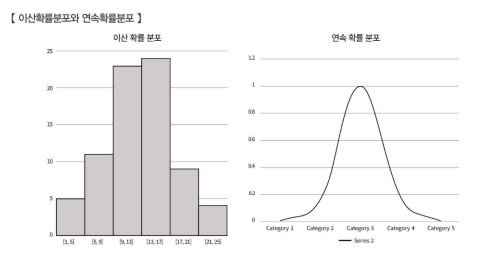
AnB = ∅

확률변수

무작위 실험을 했을 때 특정 확률로 발생하는 각각의 결과를 수치적 값으로 표현하는 변수를 확률변수라 한다 두 번 연속으로 동전 던지기 실험에서 동전의 앞면 혹은 뒷면이 나올 확률을 가지고 발생하는 결과에 앞면닝 경 우 '1', 뒷면일 경우 '0'이라는 실수값을 부여할 때, 바로 그 실수값에 부여하는 변수를 확률변수라 한다. 확률변수는 다시 변수의 특성에 따라 이산확률변수와 연속확률변수로 구분된다.

확률분포

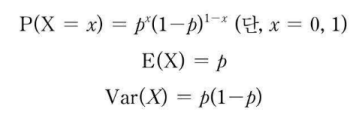
확률변수의 모든 값과 그에 대응하는 확률이 어떻게 분포하고 있는지가 바로 확률분포이다. 이때 확률변수에 의해 정의된 실수를 확률에 대응시키는 함수를 확률함수라 한다.

이산확률분포

베르누이 분포

확률변수 X가 취할 수 있는 값이 두 개인 경우로 일반적으로 한 번의 시행을 할 때 성공과 실패로 나눌 수 있는 성공 할 확률이 p인 분포를 의미

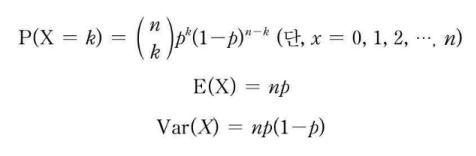
하나의 동전을 던져서 앞면이 나올 확률, 제비뽑기에서 당첨될 확률, 시험에 합격하거나 혹은 불합격할 확률 등을 예 를 들 수 있음



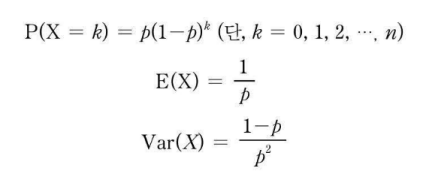
이항 분포

이항 분포는 n번의 베르누이 시행(성공 또는 실패)에서 k번 성공할 확률의 분포를 의미

하나의 동전을 3번 던져서 앞면이 2번 나올 확률, 하나의 주사위를 5번 던져서 1이 한 번 나올 확률, 3번의 제비뽑기 에서 1번 당첨될 확률 등을 예로 들 수 있음

기하 분포

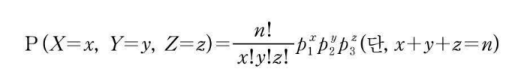
성공 확률이 p인 베르누이 시행에서 처음으로 성공이 나올 때까지 k번 실패할 확률의 분포를 의미 동전을 던져서 3번째에 앞면이 나올 확률, 주사위를 던져서 4번째에 1이 나올 확률, 제비뽑기를 복원 추출로 시행할 때 5번째에 당첨될 확률 등을 예를 들 수 있음



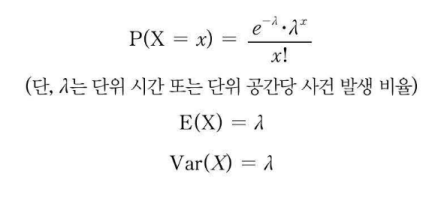
다항 분포

이항 분포를 확장한 개념으로, n번의 시행에서 각 시행이 3개 이상의 결과를 가질 수 있는 확률의 분포를 의미

주사위를 n번 던졌을 때 1의 눈이 p1의 확률로 x번, 2의 눈이 p2의 확률로 y번, 3 이상의 눈이 p3의 확률로 z번 나올 확률 등을 예로 들 수 있음

포아송 분포

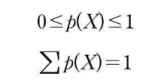
단위 시간 또는 단위 공간 내에서 발생할 수 있는 사건의 발생 횟수에 대한 확률분포를 의미 8시간 동안 3번의 장난전화가 왔을 때 1시간 동안 장난전화가 2번 올 확률, 5페이지 안에 3개의 오타가 있다면 1페 이지 안에 2개의 오타가 있을 확률 등을 예로 들 수 있음

이산확률변수

확률변수가 취할 수 있는 실수 값의 수를 셀 수 있는 변수를 이산확률변수라 한다.

이산확률변수는 셀 수 있는 실수값을 취함

서로 배반인 사건들의 합집합의 확률은 각 사건의 확률의 합이다.

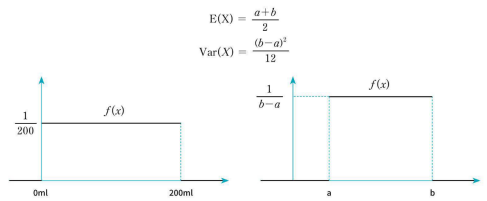


연속확률분포

균일 분포

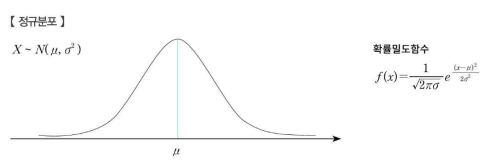
균일 분포는 연속형 확률변수인 X가 취할 수 있는 모든 값에 대하여 같은 확률을 갖고 있는 분포를 의미 얼마나 들어 있는지 모르는 200ml 우유팩 속에 들어 있는 우유의 양 등과 같은 것을 예로 들 수 있음

다음 두 개의 균일 분포 모두 그래프 아래 면적의 넓이는 확률의 총합인 1이다.

정규분포

가장 대표적인 연속형 확률분포 중 하나로 평균이 μ이고, 표준편차가 σ인 분포를 의미

한 학교의 1학년 수학 점수의 분포, 전국 남성의 키 등과 같은 것을 예로 들 수 있음

분포의 모양은 평균값에 가장 많이 집중되어 있고 평균에서 멀어질수록 빈도수가 낮은 종 모양의 그래프를 가짐 표준정규분포

정규분포는 평균μ와 표준편차 σ에 의하여 다양한 모양을 가질 수 있기 때문에 확률변수가 일정 범위 내에 포함될 확 률을 매번 계산해야 하는 번거로움이 생김

문제를 해결하기 위해서 등장한 것이 표준정규분포

표준정규분포는 평균이 0, 표준편차가 1인 정규분포를 의미

아래의 공식을 사용하여 정규분포를 따르는 확률변수 X를 표준정규분포를 따르는 확률변수 Z로 변환할 수 있음 이 작업을 표준화라고 함

표준화는 머신러닝과 딥러닝에서 중요한 도구로 사용됨

표준화된 확률분포는 표준정규분포 표를 활용하여 쉽게 확률 값을 구할 수 있음

표준정규분포의 확률밀도함수(PDF)

f(z) =1

√2πe−z22

표준화 공식

Z =X − μ

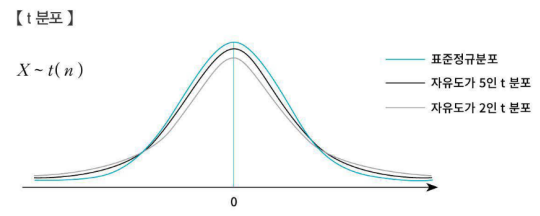
σ

**t-**분포

자유도가 n인 t분포는 표준정규분포와 마찬가지로 평균이 0이고 좌우가 대칭인 종 모양의 그래프지만 정규분포보다 두꺼운 꼬리를 가짐

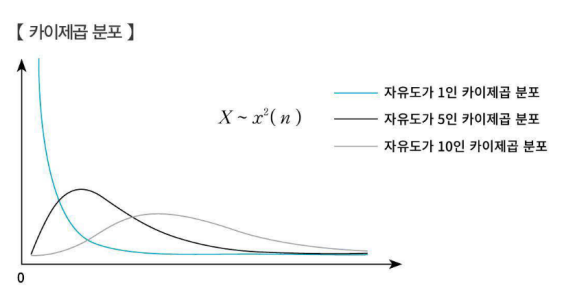
표준정규분포를 활용하여 모평균(모수)을 추정하기 위해서는 모표준편차를 사전에 알고 있어야 한다 그러나 현실적으로 모표준편차를 모르기 때문에 t 분포를 이용하여 모평균 검정 또는 두 집단의 평균이 동일한지 계 산하기 위한 검정통계량으로 활용

자유도가 커질수록 t 분포는 표준정규분포에 가까워 진다.

자유도는 표본자료들이 모집단에 대한 정보를 주는 독립적인 자료의 개수를 의미

카이제곱 분포

표준정규분포를 따르는 확률변수 Z1,Z2,Z3. . . . ,Zn의 제곱의 합 X는 자유도가 n인 카이제곱 분포를 따름 카이제곱 분포는 모평균과 모분산을 모르는 두 개 이상의 집단 간 동질성 검정 또는 모분산 검정을 위해 활용

기댓값**,** 분산**,** 표준편차

기댓값

분산

표준편차

그외 개념

첨도와 왜도

공분산

상관계수

추정과 가설검정