**2주차 스터디 노트: 확률 분포와 샘플링**

**주제: 확률 분포의 이해**

**스터디 목표**

* **확률 분포의 정의와 중요성 이해**: 확률 분포가 무엇인지, 데이터 분석에서 왜 중요한지 설명할 수 있다.
* **주요 확률 분포의 특성과 적용 사례 학습**: 정규분포, 이항분포, 포아송분포의 특성과 실제 사례를 통해 각각의 분포를 이해하고 적용할 수 있다.
* **확률 분포의 수식 이해 및 계산**: 각 확률 분포의 수식을 이해하고, 주어진 데이터에 대해 직접 계산할 수 있다.
* **데이터 시각화를 통한 분포 분석 능력 배양**: 데이터 시각화를 통해 다양한 확률 분포의 특성을 분석하고 해석할 수 있다.

**1. 확률 분포의 개요**

**확률 분포의 정의**

* **확률 분포**: 확률 변수가 가질 수 있는 모든 값들과 각 값이 나타날 확률을 나타내는 함수.
* **역할**: 데이터의 분포를 이해하고, 불확실성을 모델링하며, 예측을 위한 기초 제공.

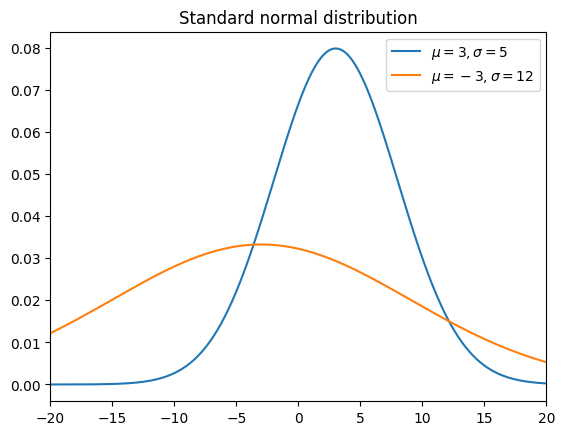
**확률 분포의 종류**

* **이산 확률 분포**: 특정 값만 가질 수 있는 확률 변수의 분포. 예: 이항분포, 포아송분포.
* **연속 확률 분포**: 연속적인 값을 가질 수 있는 확률 변수의 분포. 예: 정규분포.

**2. 정규분포(가우시안 분포)**

**정규분포의 정의와 특성**

* **정규분포 (Normal Distribution)**: 연속형 변수에 대한 확률 분포의 일종으로, 데이터가 평균을 중심으로 대칭적으로 분포.



* **특성**:
  + 종 모양의 곡선.
  + 평균과 표준편차로 정의됨.
  + 중앙에 위치한 평균 값이 가장 큰 빈도를 가짐.
  + 표준편차가 작을수록 곡선이 좁아지고, 클수록 넓어짐.

**정규분포의 수식(확률밀도함수=그래프가 나타내는 함수)**

* *μ*: 평균
* *σ*: 표준편차

**정규분포의 예시**

* 키, 몸무게, 시험 점수 등 자연현상과 관련된 많은 데이터가 정규분포를 따름.

**2-1. 표준정규분포:** 정규분포에서 평균이 0이고, 표준편차가 1일 때를 의미. 이것만으로 의미가 있지는 않고 정규화를 통해 의미를 가짐

텍스트, 그래프, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

**표준정규분포의 수식(확률밀도함수)**

**(연속형) 확률 밀도 함수(Probability Density Function, PDF) f(x)의 특징**

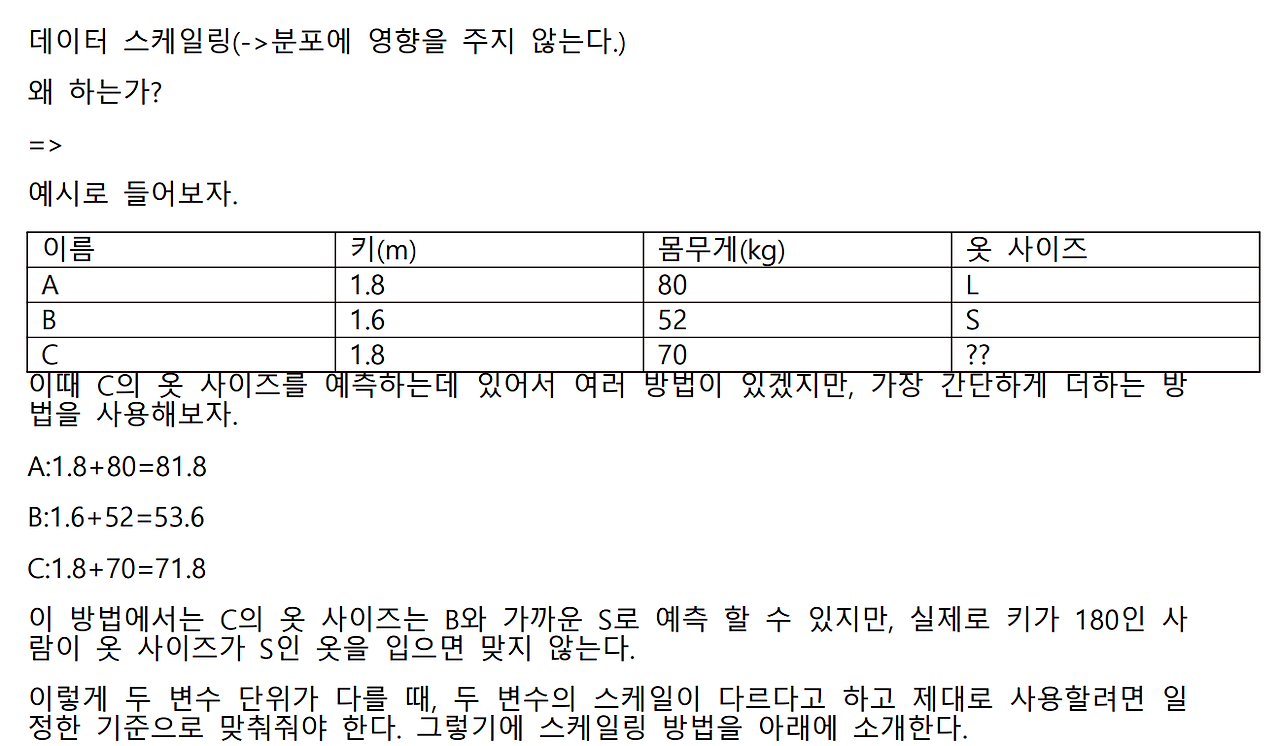
,

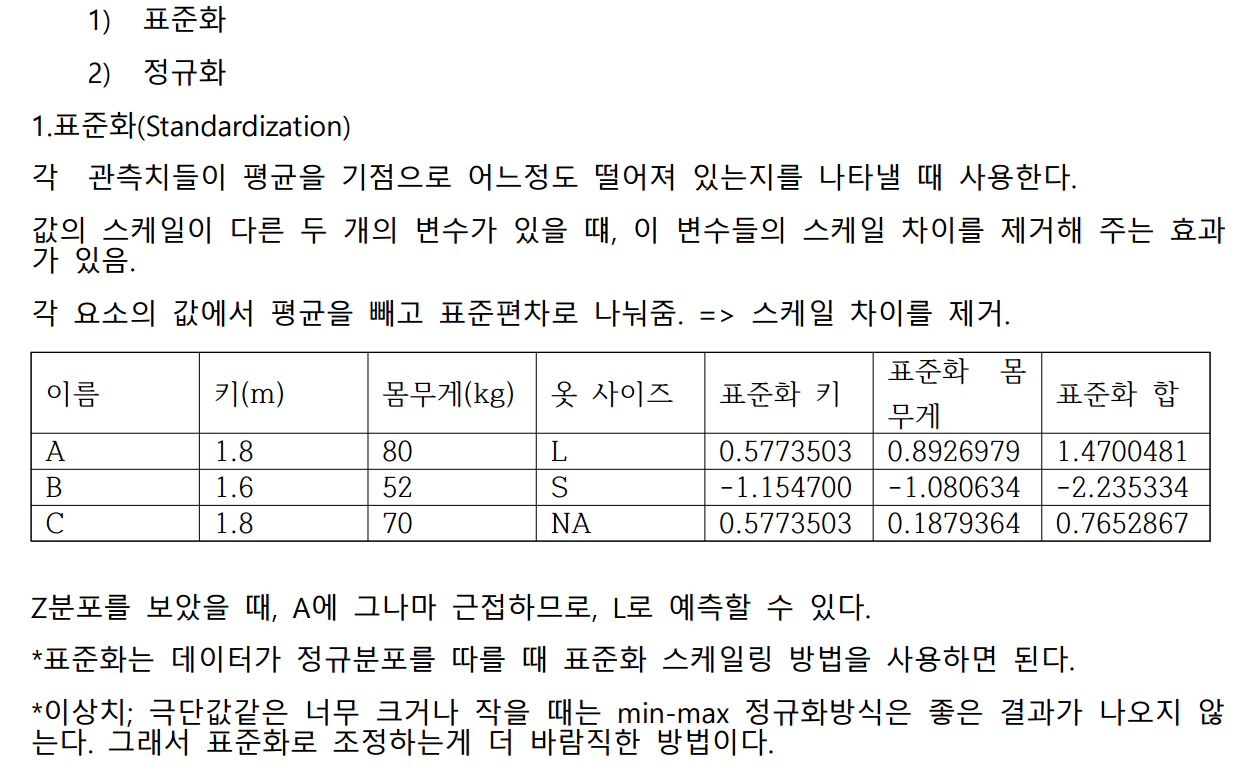
* + 전체 확률은 항상 1(100%)이므로

**(연속형)확률 분포 함수(Cumulative Distribution Function, CDF)**

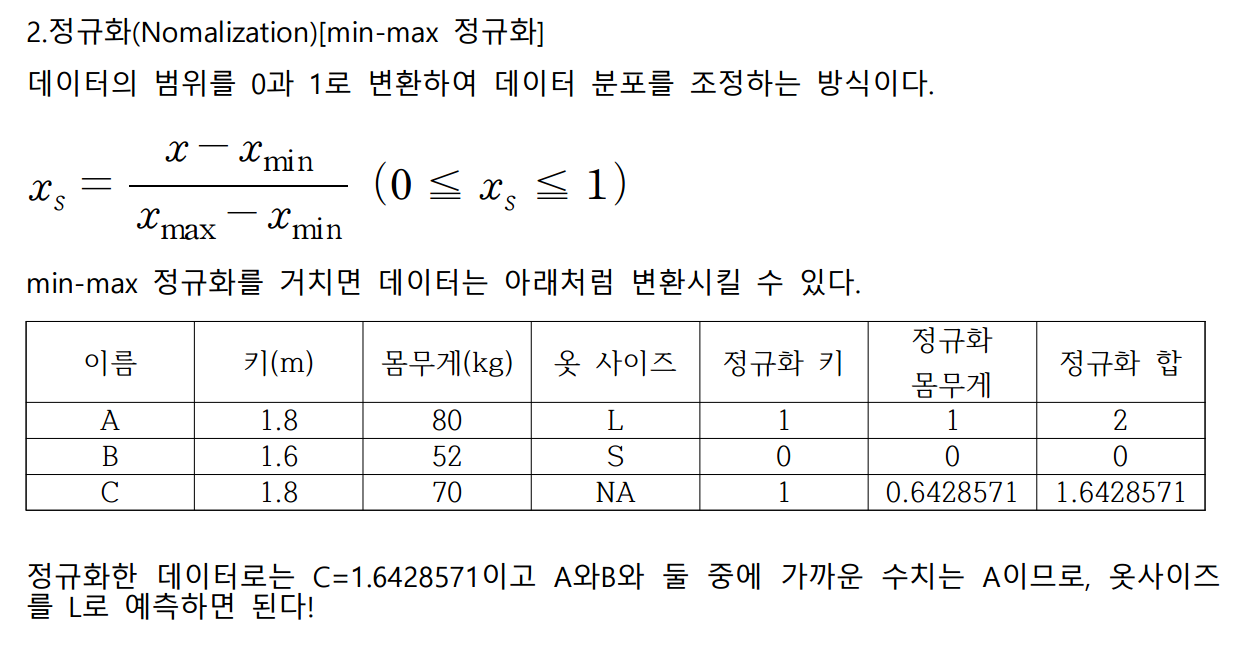
즉, CDF F(X)를 미분하면 PDF f(x)임

**2-2. 표준화(Standardization):** 정규분포의 속성을 가지도록 데이터를 평균이 0, 표준편차를 1로 바꾸는 것을 의미. 정규분포를 따르는 모든 자료에 평균을 빼고 표준편차로 나누면 됨





**2-3. 정규화(Normalization) :** 데이터의 범위를 0과 1 사이로 조정하는 것을 의미



**\*피쳐 스케일링(Feature Scaling):** 변수간 스케일의 차이가 클 때(예를 들어 연령과 아파트 가격 등) 스케일을 맞춰 상관관계의 편향을 없애줌. 스케일을 맞추지 않을 경우 값이 큰 변수인 아파트 가격이 결과에 더 큰 영향을 미칠 수 있음

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

정규화와 표준화의 차이점

**3. 이항분포**

**이항분포의 정의와 특성**

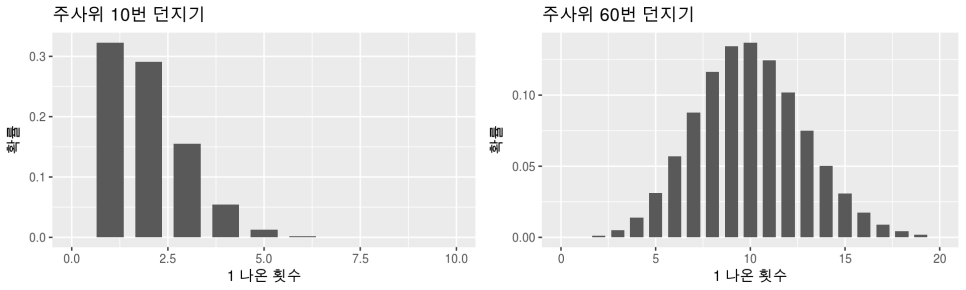
* **이항분포 (Binomial Distribution)**: 이산형 변수에 대한 확률 분포로, 고정된 횟수의 독립적인 시행에서 성공 횟수를 나타냄.
* **특성**:
  + 각 시행은 두 가지 결과 (성공 또는 실패) 중 하나만 가짐.
  + 각 시행의 성공 확률은 일정함.
  + 시행 횟수와 성공 확률로 정의됨.

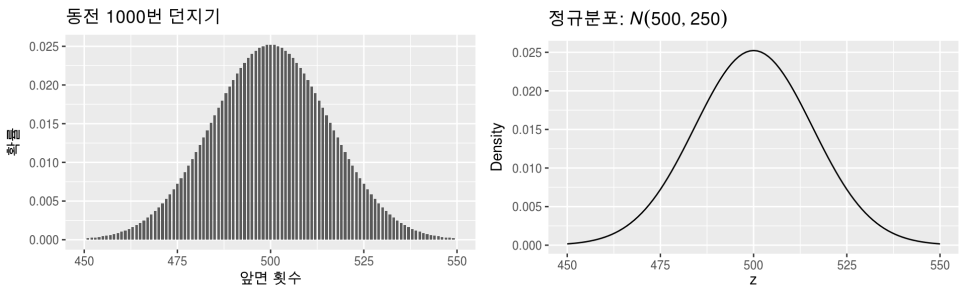
**이항분포의 수식**

* *n*: 시행 횟수
* *k*: 성공 횟수
* *p*: 성공 확률

**이항분포의 예시**

* 동전 던지기, 시험에서의 정답 개수, 제품의 불량률 등.





시행 횟수가 많아지면 사실상 이항분포도 정규분포곡선과 가까운 모양이 됨

**4. 포아송분포**

**포아송분포의 정의와 특성**

* **포아송분포 (Poisson Distribution)**: 이산 확률 분포의 일종으로, 고정된 시간 또는 공간에서 사건이 발생하는 횟수를 나타냄.
* **특성**:
  + 사건 발생이 독립적임.
  + 단위 시간 또는 단위 공간에서 사건 발생률이 일정함.
  + 사건 발생 횟수로 정의됨.

**포아송분포의 수식**

* *λ*: 단위 시간 또는 공간에서의 평균 사건 발생률
* *k*: 사건 발생 횟수

**포아송분포의 예시**

* 특정 시간 동안의 콜센터에 걸려오는 전화 수, 특정 구역 내 교통사고 발생 건수 등.

**과제와 실습**

**과제 데이터셋**

* 정규분포: 학생들의 시험 점수

**정규분포 실습**

* 주어진 시험 점수 데이터를 사용하여 정규분포를 시각화하고, 평균과 표준편차 계산