확률분포 (ProbabilityDistribution)

가장 널리 알려진 확률분포는 정규분포다. 정규분포는 특정 확률 밀도 함수로 (Normal PDF)부터 얻어진 확률들의 패턴으로 정의되는데 그 패턴이 종모양으로 중간이 높고 양 가장자리로 갈수록 낮아지는 모양이다. 연속형 확률분포의 예는 정규분포를 포함하여 일양분포, 지수분포, 로그정규분포, t분포, 카이제곱분포, f분포, 감마분포, 베타분포 등이 있다. 이산형 확률분포는 베르누이분포, 이항분포, 포아송분포, 기하분포, 초기하분포, 음이항분포 등이 대표적이다. 여기서 중요한 개념인 모수 (parameter)가 등장한다. 모든 확률분포는 한 개 이상의 모수를 가지고 있으며 이는 확률분포의 모양을 결정한다. 예를 들어 정규분포의 경우 평균과 분산, 2개의 모수가 있고 이들이 정규분포의 모양을 결정한다. 많은 경우 분포의 모수는 알려져 있지 않으며 모수를 추측해보는 과정을 통계학에서는 추정 (Estimation)이라고 한다.

모수적 모델 vs. 비모수적 모델

모수와 비모수의 차이에 대해 설명할 준비가 되었다. 모수적 모델은 알려진 확률분포 (위에서 설명한)를 기반으로 [redacted] 모수를 추정하는 과정이 포함되어 있는 모델을 통칭한다. 가장 대표적인 방법론으로는 선형회귀모델이 있다. 이해하기 쉽게 단순선형회귀모델 (Simple Linear Regression Model)을 예로 들어 보자. 단순선형회귀모델은 예측변수(X)와 반응변수(Y) 사이의 관계를 직선으로 표현하는 방법이다. 즉, 반응변수를 예측변수를 이용해 예측 (혹은 설명)하는 것인데 실제로는 두 변수 사이의 랜덤성으로 인해 예측변수만을 가지고 반응변수를 100% 예측 (혹은 설명)하기는 어렵다. 이 부족한 부분은 어쩔 수 없이 오차 (error)로 표현을 해야 한다. 선형회귀모델에서는 이 오차값들이 평균이 0이고 분산이 특정 상수값을 갖는 정규분포를 따른 다는 가정하에 수립된다. 이렇듯 선형회귀모델은 모델 구축 시 알려진 확률분포 (정규분포)를 가정하기 때문에 모수적 방법론에 속한다.