주성분분석(PCA)과 공통요인분석(CFA)

# 1. 주성분분석 (Principal Component Analysis, PCA)

주성분분석(PCA)은 데이터의 차원을 축소하거나 변수들 간의 주요 구조를 파악하기 위해 사용되는 기법입니다. PCA는 원래의 변수들을 선형 결합하여 새로운 변수(주성분)를 생성하며, 이 주성분들은 데이터의 분산을 최대한 설명할 수 있도록 설계됩니다.

## PCA의 주요 특징

- 분산 극대화: 각 주성분은 데이터의 분산을 최대한 설명하도록 계산됩니다. 첫 번째 주성분은 전체 데이터의 최대 분산을 설명하며, 두 번째 주성분은 첫 번째 주성분과 직교(orthogonal)하면서 남은 분산을 최대한 설명합니다.

- 차원 축소: 주성분은 원래의 변수를 결합하여 차원을 축소합니다. 주성분의 수는 원래 변수의 수와 같지만, 보통 몇 개의 주성분만으로도 데이터의 대부분의 정보를 설명할 수 있습니다.

- 해석의 어려움: PCA에서 생성된 주성분은 단순히 수학적 변환에 의해 도출된 것이므로, 해석이 어려울 수 있습니다. 주성분은 원래 변수의 물리적 의미를 반영하지 않기 때문입니다.

- 상관행렬 또는 공분산행렬: PCA는 상관행렬이나 공분산행렬을 기반으로 계산됩니다. 상관행렬을 사용하면 각 변수는 동일한 중요도를 가지며, 공분산행렬을 사용하면 변수의 분산 크기에 따라 가중치가 다르게 부여됩니다.

## PCA의 사용 목적

- 데이터의 차원 축소: 많은 변수로 이루어진 데이터를 몇 개의 주성분으로 축소하여 분석의 복잡성을 줄이면서 중요한 정보를 보존합니다.

- 노이즈 제거: 불필요한 변수를 제거하여 데이터의 노이즈를 줄이고, 더 명확한 분석을 가능하게 합니다.

- 시각화: 고차원 데이터를 2차원 또는 3차원으로 축소하여 시각적으로 표현할 수 있습니다.

# 2. 공통요인분석 (Common Factor Analysis, CFA)

공통요인분석(CFA)은 변수들 간의 공통된 변동성을 설명하는 잠재 요인(공통요인)을 추출하는 방법입니다. 이 방법은 변수들이 몇 개의 공통된 요인에 의해 영향을 받는다는 가정하에 수행됩니다.

## CFA의 주요 특징

- 공통성(Communality): 공통요인분석에서 중요한 개념으로, 각 변수의 변동성 중 공통된 요인에 의해 설명되는 부분을 의미합니다. 공통성은 변수의 전체 변동성에서 특정 요인들이 설명할 수 있는 비율로 나타납니다.

- 요인부하량(Factor Loadings): 각 변수와 공통요인 간의 상관관계를 나타내는 계수입니다. 요인부하량이 클수록 해당 변수는 그 요인에 의해 더 많이 설명된다는 것을 의미합니다.

- 잔차(Unique Variance): 공통요인으로 설명되지 않는 부분으로, 각 변수의 특수 요인(특수성)과 측정 오차를 포함합니다.

- 모형 적합도: 공통요인분석은 모델링 접근법을 사용하므로, 모형의 적합도를 평가할 수 있습니다. 이때 최대우도법 등의 통계적 방법을 통해 모형의 적합성을 검정할 수 있습니다.

# 3. 공통성(Communality)의 개념

공통성(Communality)은 공통요인분석에서 각 변수의 변동성 중 공통요인들에 의해 설명되는 부분을 의미합니다.

- 공통성의 값: 각 변수의 공통성 값은 0에서 1 사이의 값을 가지며, 1에 가까울수록 해당 변수의 변동성이 공통요인에 의해 잘 설명된다는 것을 의미합니다.

- 공통성 추정: 초기에는 주성분분석을 통해 공통성을 추정할 수 있으며, 반복적인 과정(예: 최대우도법)을 통해 추정치를 개선할 수 있습니다.

- 공통성의 역할: 공통성 값이 낮은 변수는 공통요인으로 설명되지 않는 부분이 많다는 것을 의미하므로, 분석에서 제거하거나 별도로 고려할 수 있습니다.

# 4. 변수가 많아질 때의 분석 결과

변수의 수가 많아지면, 주성분분석과 공통요인분석의 결과에 다양한 영향을 미칠 수 있습니다.

## 주성분분석에서:

- 차원 축소의 필요성: 변수가 많아질수록, 주성분분석을 통해 차원을 축소하는 것이 더 중요해집니다. 적은 수의 주성분으로 데이터의 대부분의 분산을 설명할 수 있다면, 분석이 더 간단해지고 해석이 용이해집니다.

- 주성분의 해석: 변수가 많아질수록, 주성분의 해석이 더 어려워질 수 있습니다. 각 주성분은 많은 변수의 선형 결합이기 때문에, 해석할 때 주의가 필요합니다.

- 노이즈와 정보 손실: 너무 많은 변수를 축소하려 할 경우, 노이즈와 유용한 정보 모두가 제거될 수 있습니다. 따라서 적절한 주성분의 수를 선택하는 것이 중요합니다.

## 공통요인분석에서:

- 공통성의 변화: 변수가 많아질수록 각 변수의 공통성이 더 분산될 수 있습니다. 공통성이 낮은 변수는 공통요인으로 설명되지 않는 특성이 많을 수 있습니다.

- 요인의 해석: 변수가 많아지면, 추출된 요인이 더 복잡해질 수 있으며, 각 요인의 해석이 더 어려워질 수 있습니다. 이는 각 요인에 관련된 변수가 많아지기 때문입니다.

- 모형의 적합성: 변수가 많아지면 모델의 적합성을 유지하는 것이 어려워질 수 있습니다. 따라서 변수의 선택에 신중해야 하며, 필요에 따라 변수의 수를 줄이는 것이 필요할 수 있습니다.

# 요약

주성분분석(PCA)은 데이터의 분산을 최대한 설명하는 주성분을 추출하여, 데이터를 축소하거나 구조를 파악하는 데 사용됩니다. 주로 데이터 차원 축소와 노이즈 제거에 유용하지만, 주성분의 해석이 어려울 수 있습니다. 공통요인분석(CFA)은 변수들 간의 공통된 변동성을 설명하는 요인을 추출하는 데 중점을 두며, 공통성(Communality)을 기반으로 분석을 수행합니다. 주로 심리 측정과 데이터 구조 분석에 사용됩니다. 공통성은 각 변수의 변동성 중 공통된 요인에 의해 설명되는 부분을 의미합니다.