# 데이터 분석 개요

## 1. 서론

21세기 들어 데이터는 '새로운 석유'라고 불릴 정도로 가치 있는 자원으로 인식되고 있다. 기업, 공공기관, 연구기관은 방대한 양의 데이터를 수집하고 이를 분석하여 경영 의사결정, 정책 수립, 연구개발에 활용한다.
데이터 분석(Data Analysis)이란 단순히 데이터를 모으는 것을 넘어, 데이터를 정제하고, 패턴을 발견하며, 미래를 예측하고, 문제 해결을 위한 근거를 도출하는 과정 전체를 의미한다.

데이터 분석은 크게 설명적 분석(Descriptive), 진단적 분석(Diagnostic), 예측적 분석(Predictive), 처방적 분석(Prescriptive) 네 가지 유형으로 구분된다. 설명적 분석은 과거에 무슨 일이 있었는지를 보여주고, 진단적 분석은 왜 그런 일이 일어났는지를 규명한다. 예측적 분석은 앞으로 어떤 일이 일어날지를 전망하고, 처방적 분석은 무엇을 해야 하는지 대안을 제시한다.

오늘날 데이터 분석은 인공지능(AI), 사물인터넷(IoT), 스마트팩토리 등과 결합하여 산업 전반의 혁신을 이끌고 있다. 따라서 데이터 분석에 대한 이해는 모든 산업 종사자에게 필수 역량으로 자리 잡고 있다.

## 2. 데이터 분석의 단계

데이터 분석은 일련의 절차를 통해 진행된다. 일반적으로 다음의 단계를 따른다.

### (1) 문제 정의

분석은 문제 인식에서 출발한다. "고객 이탈률이 왜 높은가?", "생산 라인의 불량률을 줄일 방법은 무엇인가?" 등 명확한 문제 정의 없이는 데이터 분석의 방향성이 흐려진다.

### (2) 데이터 수집

데이터는 내부 시스템(ERP, MES, CRM 등)과 외부 데이터(공공 데이터, 웹 크롤링, 시장 조사 등)에서 수집된다. 최근에는 센서 데이터, 로그 데이터, 소셜 미디어 데이터 등 비정형 데이터의 비중이 급격히 증가하고 있다.

### (3) 데이터 정제 및 전처리

현장에서 수집된 데이터는 결측치, 이상치, 중복이 존재한다. 따라서 분석 전에 데이터 정제(Cleaning)와 전처리(Preprocessing)가 필수적이다. 예를 들어 결측치는 평균 대체, 삭제, 추정으로 처리할 수 있고, 이상치는 IQR, Z-score 기법으로 탐지할 수 있다.

(4) 탐색적 데이터 분석(EDA)

EDA는 데이터의 분포, 특성, 변수 간 관계를 시각적으로 탐색하는 단계다. 히스토그램, 박스플롯, 산점도, 상관계수 등이 사용된다. 이 단계는 분석가가 데이터에 대한 직관을 얻고, 이후 모델링 방향을 결정하는 중요한 역할을 한다.

(5) 통계적 분석 및 모델링

데이터 분석의 핵심 단계다. 통계학적 기법(회귀분석, 분산분석, 가설검정)과 머신러닝 기법(분류, 군집, 예측모델)을 활용하여 문제를 해결한다. 제조업에서는 불량률 예측, 금융업에서는 신용평가, 유통업에서는 수요 예측에 활용된다.

(6) 결과 해석 및 시각화

모델의 결과를 단순히 수치로 제시하는 것이 아니라, 의사결정자가 이해할 수 있도록 스토리텔링과 시각화를 활용한다. 예컨대, 대시보드 형태의 시각화는 경영자가 실시간으로 의사결정을 내리는 데 도움을 준다.

(7) 실행 및 모니터링

데이터 분석 결과는 정책/전략으로 실행되며, 주기적인 모니터링을 통해 피드백이 이루어진다. 분석은 일회성 작업이 아니라 지속적인 과정이다.

3. 데이터 분석 기법
(1) 기술통계(Descriptive Statistics)

평균, 중앙값, 분산, 표준편차, 상관계수 등 데이터의 특성을 요약한다. 예를 들어 제조 공정에서 하루 평균 생산량, 불량률 분산을 계산하면 공정 안정성을 파악할 수 있다.

(2) 추론통계(Inferential Statistics)

표본 데이터를 통해 모집단을 추정하고 가설을 검정한다. 예를 들어, 100개 제품을 샘플링하여 불량률이 5% 이하라는 가설을 검정할 수 있다.

(3) 예측모델링(Predictive Modeling)

회귀분석, 시계열 분석, 머신러닝 알고리즘을 활용하여 미래를 예측한다. 판매량 예측, 수요 전망, 주가 예측 등이 대표적이다.

(4) 데이터 마이닝 & 머신러닝

분류(Classification): 이메일이 스팸인지 아닌지 분류
군집화(Clustering): 고객을 구매 패턴에 따라 그룹화
연관규칙(Association Rule): 장바구니 분석(맥주와 기저귀)
시계열(Time Series): 시점별 데이터 패턴과 추세 분석

(5) 텍스트 마이닝 / 자연어 처리

리뷰, SNS, 보고서 등 비정형 텍스트 데이터를 분석하여 감성 분석, 키워드 추출, 주제 모델링 등을 수행한다.

4. 데이터 분석 도구와 환경
(1) 프로그래밍 언어

Python: Pandas, Numpy, Matplotlib, Scikit-learn → 데이터 전처리·모델링·시각화 전반

R: 통계학 기반 분석과 시각화에 강점

(2) 데이터베이스와 SQL

대규모 데이터를 관리하기 위해 SQL, NoSQL이 활용된다.
PLSQL, HiveQL 등은 기업 내 빅데이터 환경에서 중요하다.

(3) 시각화 도구

Tableau, Power BI: 비전문가도 쉽게 데이터 시각화 가능
Python Seaborn, Plotly: 프로그래밍 기반 시각화

(4) 클라우드 및 빅데이터 플랫폼

AWS, Azure, GCP의 데이터 분석 플랫폼
Hadoop, Spark는 대용량 데이터 분산 처리에 활용된다.

5. 데이터 분석의 산업 활용 사례
(1) 제조업

스마트팩토리: 센서 데이터 기반 설비 고장 예측
품질관리: 불량률 패턴 분석 및 공정 개선

(2) 유통/마케팅

고객 구매 패턴 분석 → 맞춤형 프로모션
상품 진열 최적화, 가격 탄력성 분석

(3) 금융

신용평가 모델 → 대출 승인 여부 판단
이상거래 탐지(Fraud Detection)

(4) 의료/헬스케어

환자 진단 데이터 분석 → 질병 예측
유전자 데이터 분석 → 맞춤형 치료

(5) 공공 분야

교통 데이터 분석 → 신호 최적화, 혼잡도 예측
인구통계 분석 → 정책 수립 근거 제공

6. 데이터 분석의 한계와 고려사항

데이터 품질 문제: Garbage In, Garbage Out → 잘못된 데이터는 잘못된 결과 초래
해석의 오류: 상관관계 ≠ 인과관계
윤리적 문제: 개인정보 보호, 알고리즘 편향
조직적 한계: 분석 결과가 실제 의사결정으로 연결되지 않는 경우

## 7. 결론

데이터 분석은 단순한 기술이 아니라 문제 해결 방법론이다. 분석가는 데이터의 흐름을 이해하고, 적절한 통계적 기법과 도구를 선택하여 문제 해결에 기여해야 한다.
오늘날 기업은 데이터 분석을 통해 경쟁우위 확보, 비용 절감, 품질 향상을 이루며, 공공 부문은 정책 효율성과 국민 서비스 향상을 기대할 수 있다.

향후 인공지능 기술과 융합된 데이터 분석은 더욱 정교해질 것이다. 따라서 분석가는 통계학적 사고, 프로그래밍 역량, 도메인 지식을 고루 갖추어야 하며, 이를 통해 데이터 기반의 미래 사회에서 핵심적인 역할을 하게 될 것이다.

# 제조데이터 분석

□ 공급사(A사)의 관점에서 필요한 분석

1. 기본 수요 분석

- 파트별 총 수요량
  · D일, D+1 … 하루 총 필요 수량
  · D+2, D+3, D+4 … 부족분 수요 추세
  · D+5 ~ D+30 … 미래 일자별 수요 추세
  · D+31 ~ D+45 … 미래 15일간 수요 추세

- 시간대별 수요 집중도

하루 중 언제 가장 많이 필요한지 (예: 오전 집중, 야간 분산)

→ 분석 목적: 생산/출하 타이밍 맞추기

2. 부족분(과부족) 분석

- D+2일 과부족수량 컬럼 활용
  · 어떤 파트에서 반복적으로 부족이 발생하는지
  · 특정 시간대·일자에 공급 지연 리스크 있는지

→ 분석 목적: 공급 우선순위 정하기

3. 변동성·안정성 분석

- 수요 변동 계수(CV = 표준편차/평균) 계산
  · 변동성이 큰 파트 = 재고 안전재고 필요
  · 안정적인 파트 = JIT(Just-In-Time) 가능

→ 분석 목적: 파트별 공급전략 차별화

4. 예측 분석

- 과거 패턴을 기반으로 향후 1~2주 수요 예측
  “특정 파트는 주말에 급증”, “월초에 집중” 같은 패턴 찾기

→ 분석 목적: 생산·발주 선행 계획

5. 리스크 관리

- A사가 공급하는 1~10번 파트에 대해:
  · 누적 부족분 Top N 파트 파악
  · 리드타임 고려 시 공급 못 맞출 위험이 있는지
  · 고객사와 SLA(Service Level Agreement) 기준 충족 여부

→ 분석 목적: 계약 이행률/고객 만족도 관리

■ A사가 해야 할 대표적 분석 예시

- 파트별 일일 총 수요량 (Part 1~10 합산)
→ “하루에 최소 몇 개를 생산해야 하는가?”

- 시간대별 수요 집중도 분석
→ “몇 시에 출하를 맞춰야 하는가?”

- D+N일 수요 예측 그래프
→ “앞으로 생산능력을 어떻게 배분할 것인가?”

- 부족분 히트맵
→ “리스크가 가장 큰 시간대/파트는 어디인가?”

√ 정리
즉, 공급사 A사는 단순히 “수요량 총합”만 보는 게 아니라,
(1) 총 수요 + (2) 부족분 리스크 + (3) 변동성 + (4) 예측을 분석해서
→ 생산계획(얼마나 만들지) + 공급계획(언제 납품할지)을 결정해야 합니다.

# 기초 통계 개념 10가지

1. 모집단과 표본 (Population & Sample)

- 모집단: 연구 대상 전체 집합
- 표본: 모집단에서 추출한 일부 데이터
  예: 모든 고객(모집단) 중 설문에 응답한 1,000명(표본)

2. 평균 (Mean)

- 데이터 값들의 산술적 평균
  예: 학생들의 수학 점수 평균이 75점

3. 중앙값 (Median)

- 데이터를 크기순으로 정렬했을 때 중앙에 위치한 값
  예: 소득 분포에서 평균보다 중앙값이 현실적인 “중간 소득” 지표로 활용

4. 분산과 표준편차 (Variance & Standard Deviation)

- 분산: 데이터가 평균에서 얼마나 퍼져 있는지
- 표준편차: 분산의 제곱근, 데이터 흩어짐 정도
  예: 시험 점수가 모두 비슷하면 표준편차 ↓, 들쭉날쭉하면 ↑

5. 확률 (Probability)

- 사건이 일어날 가능성을 0~1 사이 숫자로 표현
  예: 주사위를 던져 6이 나올 확률 = 1/6

6. 확률분포 (Probability Distribution)

- 확률변수가 취할 수 있는 값들의 분포
  예: 시험 점수가 정규분포(종 모양)로 분포

7. 상관관계 (Correlation)

- 두 변수 간의 선형적 관계 강도 (상관계수 -1~1)
  예: 공부 시간과 시험 점수는 양의 상관관계

8. 회귀분석 (Regression Analysis)
 - 변수 간 인과적 관계를 수학적으로 모델링
   예: 광고비(독립변수)가 매출(종속변수)에 미치는 영향

9. 가설검정 (Hypothesis Testing)
 - 표본 데이터를 이용해 모집단에 대한 가설을 검증
   예: “새 교육 프로그램이 성적 향상에 효과가 있다” 검정

10. 신뢰구간 (Confidence Interval)
 - 모집단 모수를 포함할 가능성이 높은 구간
   예: “평균 만족도는 3.5~4.2 사이일 확률이 95%”