

데이터 분석의 이해2

허 경 욱

Contents

- 01 데이터 분석 기초
- 02 데이터 도구 Excel
- 03 데이터 준비
- 04 데이터 전처리
- 05 데이터 모델링
- 06 데이터 요약 분석
- 07 데이터 탐색 분석
- 08 데이터 예측 분석
- 09 데이터시각화
- 10 다양한 데이터 도구

데이터 탐색 및 예 측 분석



데이터를 설명하는 통계 기초






Excel 데이터 분석 기능 활용



데이터 예측을 위한 회귀직선과
회귀식



학습 목표:

-  데이터의 대표값의 개념을 이해하고 표시할 수 있다.
-  데이터의 분포를 이해하고 표현할 수 있다.
-  상관도 계산을 위한 함수를 이용하여 데이터의 독립변수와 종속 변수에 대한 상관관계를 수식으로 표현할 수 있다.



데이터 탐색적 분석

- 데이터의 탐색적 자료 분석 (Exploratory Data Analysis: EDA)
- 데이터의 속성 및 특징 이해
- 데이터의 패턴 파악 및 잘못된 자료들을 탐색
- 데이터 변수들간의 관계 파악
- 데이터를 분석의 가설을 정형화하고 세분화하고 분석의 방향 및 방법 제시
- 본격적인 데이터 분석에 앞서 데이터의 주요 특성을 요약 또는 시각화
- EDA는 본격적인 데이터 분석 모델링을 하기 전 데이터 이해를 돕기 위해 실시
- EDA를 위해 통계 기법을 사용

통계(統計, Statistics)

- 표준국어대사전

- 한데 몰아서 어림잡아 계산함
- 어떤 현상을 종합적으로 한눈에 알아보기 쉽게 일정한 체계에 따라 숫자로 나타냄
- (수학) 집단적 현상이나 수집된 자료의 내용에 관한 수량적인 기술. 대상이 되는 집단을 일정한 시점에서 파악하는 것을 정태 통계, 일정한 기간에서 파악하는 것을 동태 통계라 하며, 사회나 자연 현상을 정리·분석하는 수단으로 쓰기도 함

- 통계학

- 표준국어대사전: 현상을 통계에 의하여 관찰·연구하는 학문
- 불확실성에 대한 논리를 부여하는 학문, 경험과학의 한 분야이자 대부분 학문의 기초
- 다양한 정의가 존재하고 축약하면 자료를 연구하는 학문, 데이터를 분석하는 학문

- 기술통계학 : 데이터들을 수집, 정리, 요약

- 추론통계학 : 표본 자료에서 얻은 정보를 이용하여 전체 집단(단위)에 대한 정보 및 불확실한 사실에 대해 예측하는 방법과 이론을 제시

기술통계(Descriptive Statistics) ※ “Descriptive : 묘사하는, 그래서 설명하는”

- 수집한 데이터를 요약 묘사 설명하는 통계 기법
- 기술통계 기법
 - 수집한 데이터를 대표하는 값이 무엇인지 찾기 → 대표값
 - 수집한 데이터가 어떻게 퍼져 있는지를 설명하기 → 데이터 분포
- 대표값
 - 주어진 자료를 대표하는 특정 값
 - 대표값은 자료의 중심적인 경향이나 자료분포의 중심의 위치를 나타냄
 - 평균 / 중앙값 / 최빈값 등
- 데이터 분포
 - 주어진 자료의 퍼짐 정도를 표현하는 값 또는 시각화 차트
 - 분산 / 표준편차 / 사분위값 / 왜도 / 첨도 / 도수분포표 / 히스토그램

대표값

● 평균

■ 주로 산술평균을 사용

- 산술평균: 데이터 모음의 값을 모두 더한 후 데이터 개수로 나눈 값
- 산술평균외에 용도에 따라 기하평균, 조화평균, 가중평균 등을 사용하기도 함

■ 데이터 중 극단적인 값 또는 이상한 값이 섞여 있거나 최대값이나 최소값의 크기가 크게 차이 나면 평균값이 데이터 모음의 대표값으로 왜곡된 의미를 가질 수 있음

● 중앙값

■ 수집된 데이터 모음의 값을 크기별로 나열하였을 때 가장 중앙에 위치한 값

- 데이터 개수가 짝수개이면 중앙 위치의 두 데이터의 평균으로 표시하기도 함

● 최빈값

■ 수집된 데이터 모음에서 가장 자주 발생한 값(빈도수가 가장 큰 값)

- 주로 대소관계가 의미 없는 자료(명목형)에서 많이 사용됨

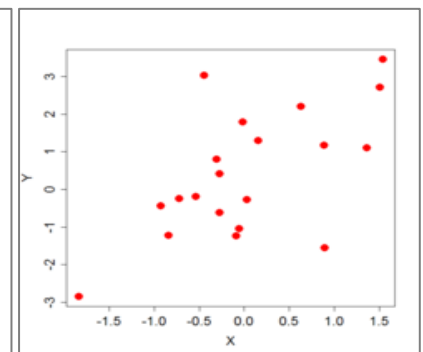
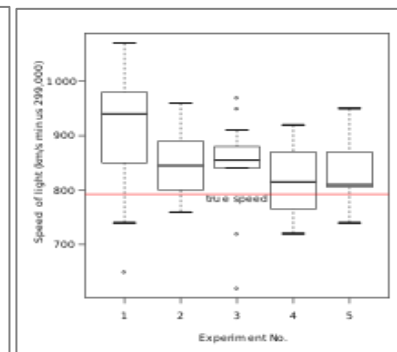
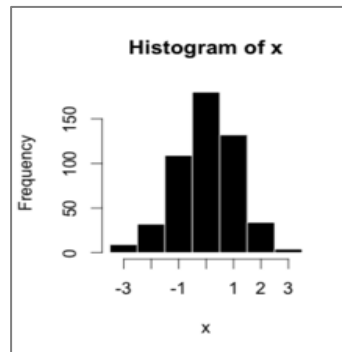
탐색적 자료 분석(EDA) 방식

● 수치적인 탐색 : 기술통계

- 평균(Mean), 최대값(Max), 최소값(Min), 중앙값(Median), 최빈값(Mode)
- 표준편차(Standard Deviation), 분산(Variance)
- 사분위수범위(Interquartile Range)
- 첨도(Kurtosis), 왜도(Skewness)

● 시각화(그래프) 탐색

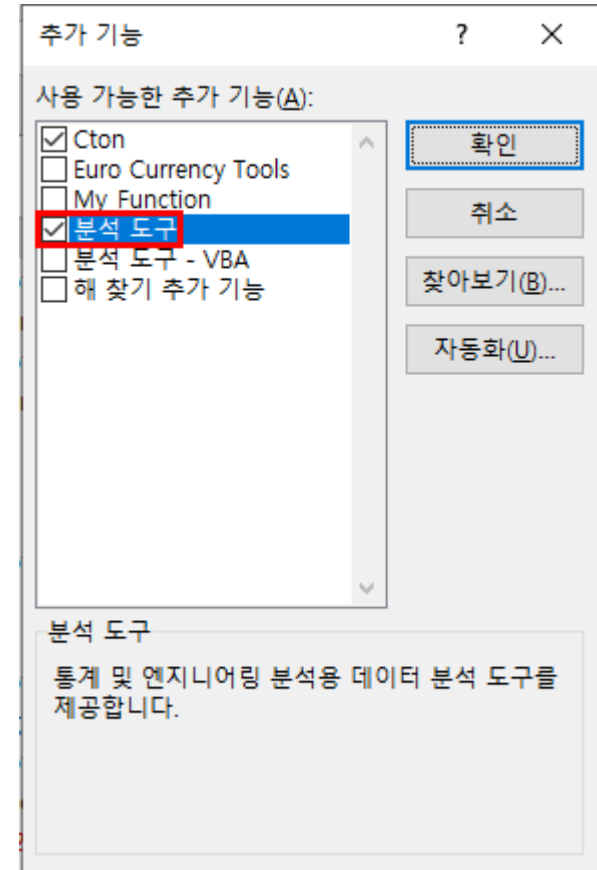
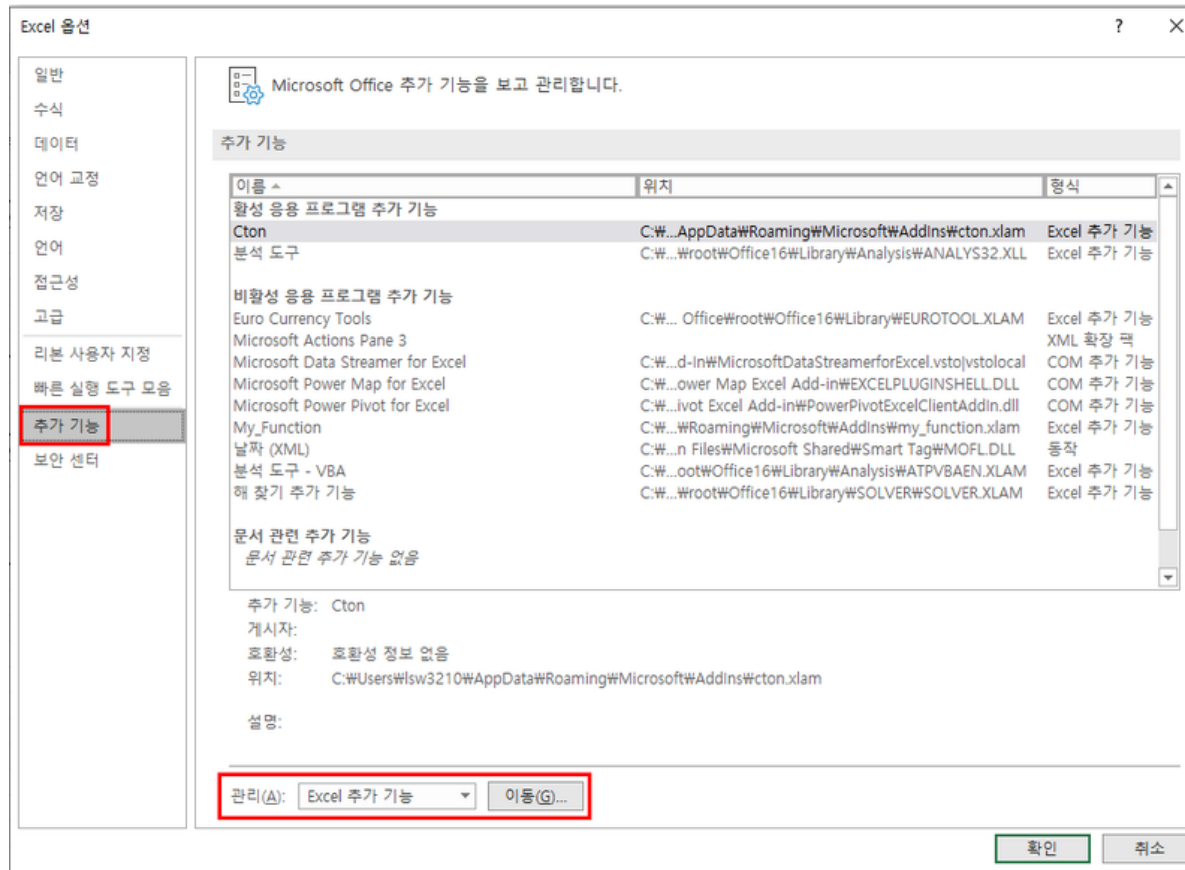
- 히스토그램(Histogram)
- 상자수염그림(Box plots)
- 산점도(Scatter plots)



데이터 탐색 분석을 위한 Excel 데이터 분석 기능

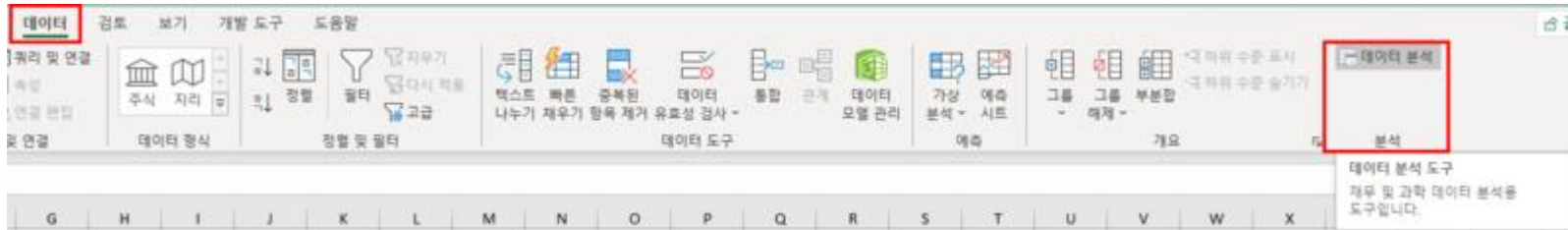
● Excel 추가 기능 → 데이터 분석 도구 설치

■ [파일] → [옵션] → [추가기능] → [엑셀 추가 기능] 이동

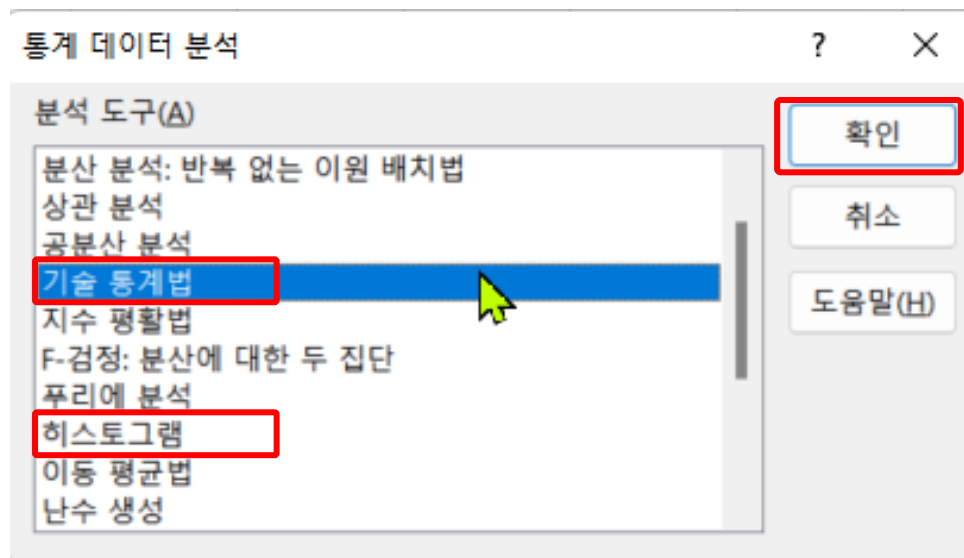


데이터 탐색 분석을 위한 Excel 데이터 분석 기능

- [데이터]탭 → 추가된 [데이터 분석] 메뉴



- 기술 통계법 / 히스토그램



기술 통계 실습

● [데이터]메뉴탭 → [데이터 분석]메뉴

■ 기술통계법

The screenshot shows the '기술 통계법' (Data Analysis) dialog box with the following settings:

- 입력: 입력 범위(I): $\$B\$2:\$B\1127 (highlighted with a red box)
- 데이터 방향: ☒ 열(C) (highlighted with a red box)
- ☒ 첫째 행 이표표 사용(L) (highlighted with a red box)
- 출력 옵션:
 - ☒ 출력 범위(O): $\$F\2 (highlighted with a red box)
 - ☒ 요약 통계량(S) (highlighted with a red box)
 - ☒ 평균에 대한 신뢰 수준(N): 95 % (highlighted with a red box)
 - ☐ K번째 큰 값(A): 1
 - ☐ K번째 작은 값(M): 1

The background shows a data table with columns A and B, and a summary table with columns E and F.

순번	점수
1	47
2	52
3	57
4	58
5	60
6	61
7	61
8	62
9	62
10	63
11	64
12	65
13	65
14	66
15	66
16	66
17	66
18	66
19	67
20	67
21	68
22	68
23	68
24	68
25	68
26	68
27	68
28	68
29	68
30	68

점수	
평균	99.617778
표준 오차	0.4511132
중앙값	100
최빈값	98
표준 편차	15.130795
분산	228.94096
첨도	-0.002133
왜도	-0.08501
범위	106
최소값	47
최대값	153
합	112070
관측수	1125
신뢰 수준(95%)	0.8851187

- 중심성
 - ✓ 중앙값은 100이고 최빈값이 98로 100과 98 사이의 값에 집중화 경향을 보임
- 변동성
 - ✓ 최대값이 153이고, 최솟값은 47로 범위가 106인 것을 알 수 있다. 또한 사람들 간의 점수 편차는 15.13 정도로 차이가 남
- 정규성
 - ✓ 왜도값은 -0.085 로 0보다 작기 때문에 부적비대칭으로 좌측으로 긴 꼬리를 가진 형태임
 - ✓ 또한 첨도는 -0.002 로 0보다 작기 때문에 가운데에 데이터들이 정규 분포보다 적게 몰려있음

기술통계 값의 의미

- 중심성 : 데이터가 어느 부분에 집중되는가?
 - 평균, 중앙값, 최빈값
- 변동성 : 데이터의 중심으로부터 얼마나 떨어져있는가?
 - 분산, 표준편차, 범위(최대값-최소값), 사분위값
- 정규성: 데이터의 분포 모양이 정규 분포인가?
 - 왜도, 첨도

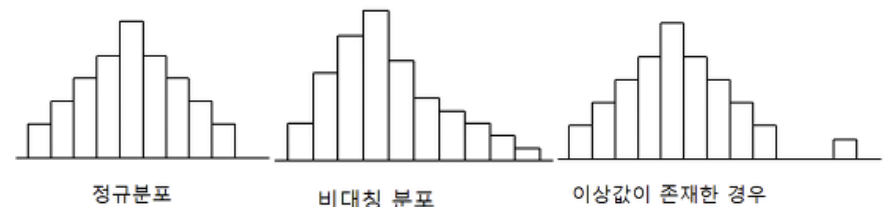
데이터 분포

● 도수분포표(Frequency table)

- 자료의 분포를 구간으로 나누고 각 구간에 속하는 값이 몇 개인지 빈도를 나타낸 표
- 전체 분포를 요약해서 파악할 수 있음
 - 자료의 개수, 자료 중 최대값과 최소값을 구함
 - 구간(계급수)의 개수를 결정: 자료의 개수나 분포에 따라 다름
 - 각 구간에 5개 이상의 값이 들어가는 것을 추천 (일반적으로 5~15구간 추천)
 - 구간의 폭(계급의 폭)을 구함: $\text{구간폭} = (\text{최대값} - \text{최소값}) / \text{구간수}$
 - 정수, 짝수, 5의배수 등의 사용을 추천
 - 구간의 경계값을 구함: 최소값에서 부터 구간폭을 더해서 구간의 경계값을 구함
 - 구간별 값의 개수(도수)를 표시

● 히스토그램(Histogram)

- 도수분포를 막대그래프로 시각화한 차트
- 그래프 모양(종모양, 비대칭 종모양 등)으로 데이터 분포를 파악할 수 있음



히스토그램(도수분포표)

● 계급을 위한 구간수와 구간크기 구하기

■ 구간수

■ 구간크기 = (최대값-최소값) / 구간수

■ 계급 경계값: 최소값부터 구간크기를 순서대로 더하여 구함

■ 계급 구간 지정

히스토그램

?

×

입력

입력 범위(I):

\$B\$2:\$B\$1127

↑

계급 구간(B):

\$H\$6:\$H\$18

↑

확인

취소

도움말(H)

☒ 이률표(L)

출력 옵션

☒ 출력 범위(O):

\$J\$6

↑

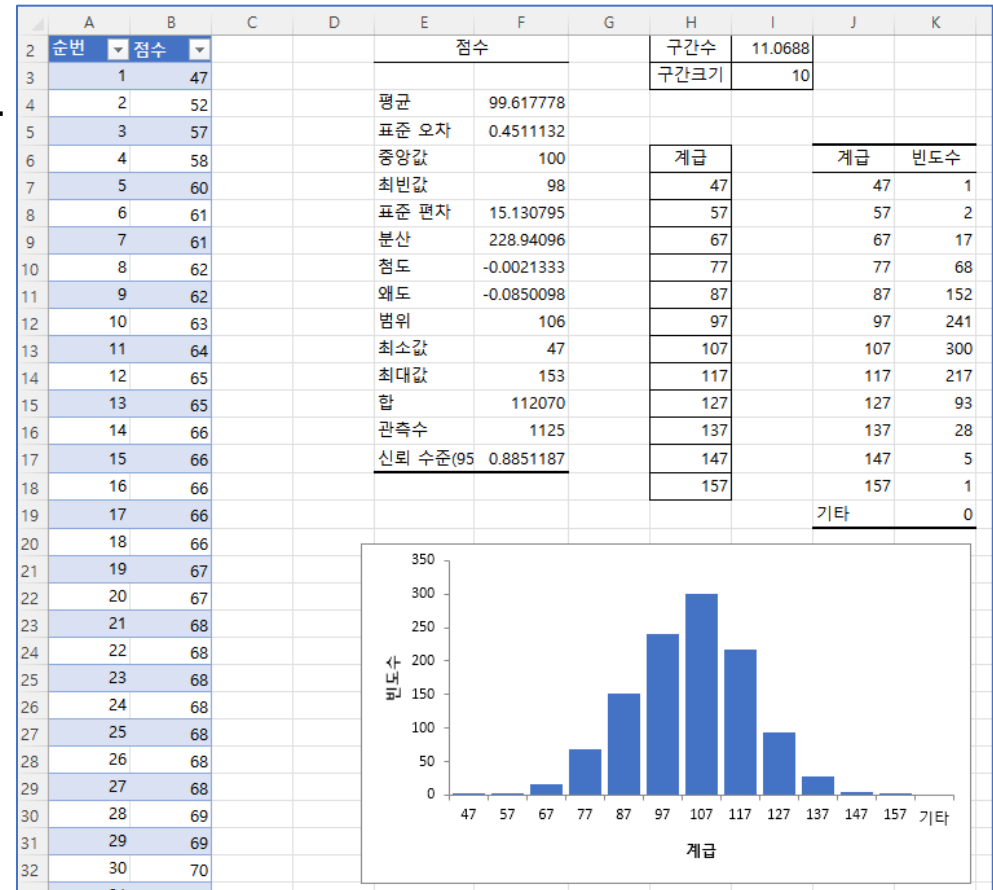
☐ 새로운 워크시트(P):

☐ 새로운 통합 문서(W)

☐ 파레토: 순차적 히스토그램(A)

☐ 누적 백분율(M)

☒ 차트 출력(C)



예측 분석이란?

- 현재 및 과거 데이터를 분석하여 미래 이벤트를 예측하는 분석 방법
- 통계 모델링, 데이터 마이닝, 머신러닝(인공지능)등의 분석 기술 사용
- 예측 분석은 이전 데이터, 통계 알고리즘, 예측 모델링, 빅데이터, 머신러닝 기술을 사용하여 향후 결과를 더 정교하게 예측하고 미래의 기회를 찾는 데 도움
- 예측 분석을 위한 다양한 빅데이터 인공지능 플랫폼 및 소프트웨어 연구 개발
- 데이터분석 기초 단계 SW인 Excel의 예측 분석 도구
 - **수식 결과의 가상 분석 : 목표값 찾기, 해 찾기, 시나리오**
 - 선형회귀관련 함수 : FORECAST(), LINEST(), TREND() 등
 - 데이터분석 메뉴 : 상관분석, 회귀분석
 - **차트 : 분산차트의 추세선 회귀식 표시**

데이터와 변수

- 분석을 위한 재료인 데이터를 수집할 때 개체, 요인, 변수로 구분할 수 있음
- 개체는 연구 대상, 요인은 특성, 변수는 요인을 구성하는 요소를 의미
 - 예를 들면 개체: 일자리 창출 사업 참여 청년
 - 요인: 청년의 개인적 특성
 - 변수: 성별, 나이, 전공, 학력, 거주지 등
- 변수(속성)중 변수들 사이에 상관관계가 존재할 수 있음
 - 영향을 주는 변수와 영향을 받는 변수가 존재할 수 있음
- 변수들의 상관관계를 분석하여 연관성을 파악
- 두 변수 사이에 존재하는 상호 의존 관계를 함수 관계로 표현하여 변수의 값이 주어지면 다른 변수의 값이 어떻게 변화하는지 예측하기도 함

독립변수와 종속변수

- 독립변수(=설명변수)

- 다른 변수의 변화를 가져오거나 영향을 미치는 원인 변수로서, 결과를 예측하게 하거나 차이를 설명하기 위해 사용되는 변수
- 실험 및 조사 등에서 연구자나 조사자에 의해 조정되는 실험 변수

- 종속변수(=결과변수)

- 독립변수의 영향으로 나타나는 결과가 되는 변수, 독립변수의 변화에 따라 변경되는 변수
- 실험 및 조사등의 결과 변수

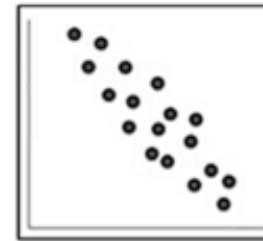
상관분석과 회귀분석

● 상관분석

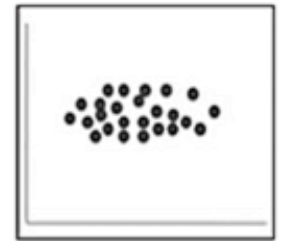
- 상관관계는 두 변수 사이의 연관성 또는 연관성을 알 수있게 해주는 분석
- 상관계수를 구하여 상관도를 나타냄
- 상관관계를 산점도로 표현할 수 있음



0<상관계수<=1



-1>=상관계수>0



상관계수=0

● 회귀분석(단순회귀분석)

- 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석 방법
- 두 변수중 독립 변수('x')의 알려진 값을 기반으로 종속 변수('y')의 값을 예측
- 선형회귀식 : $y = a + bx$ (a 는 절편, b 는 기울기)
- 결정계수 (R^2) : 독립변수가 종속변수를 얼마나 잘 설명하는가를 나타내는 수치
 - 상관계수의 제곱 : 0 ~ 1 사이의 값으로 1에 가까울 수록 잘 설명하고 있다하고 해석함

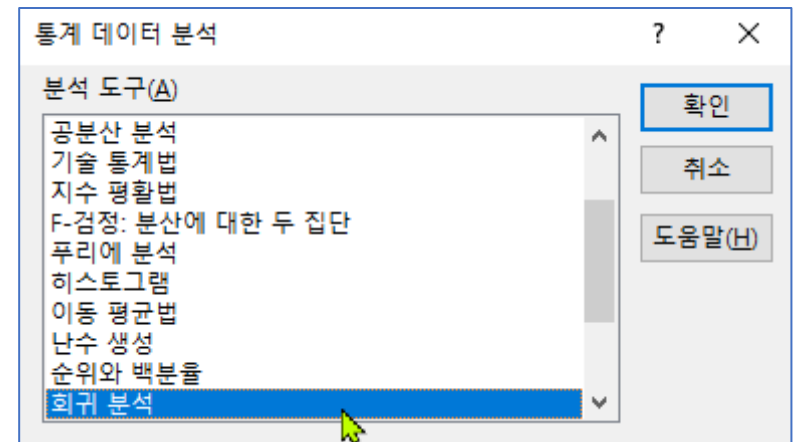
Excel의 상관관계 및 회귀분석

- 분석할 데이터는 반드시 연속된 '숫자'이어야 함
- 적절한 가설(설명변수로 결과변수를 설명하는 가설)을 수립
- 상관계수의 높음이 두 변수의 인과관계가 있다고 해석하면 안됨

상관계수별 두 변수간의 관계

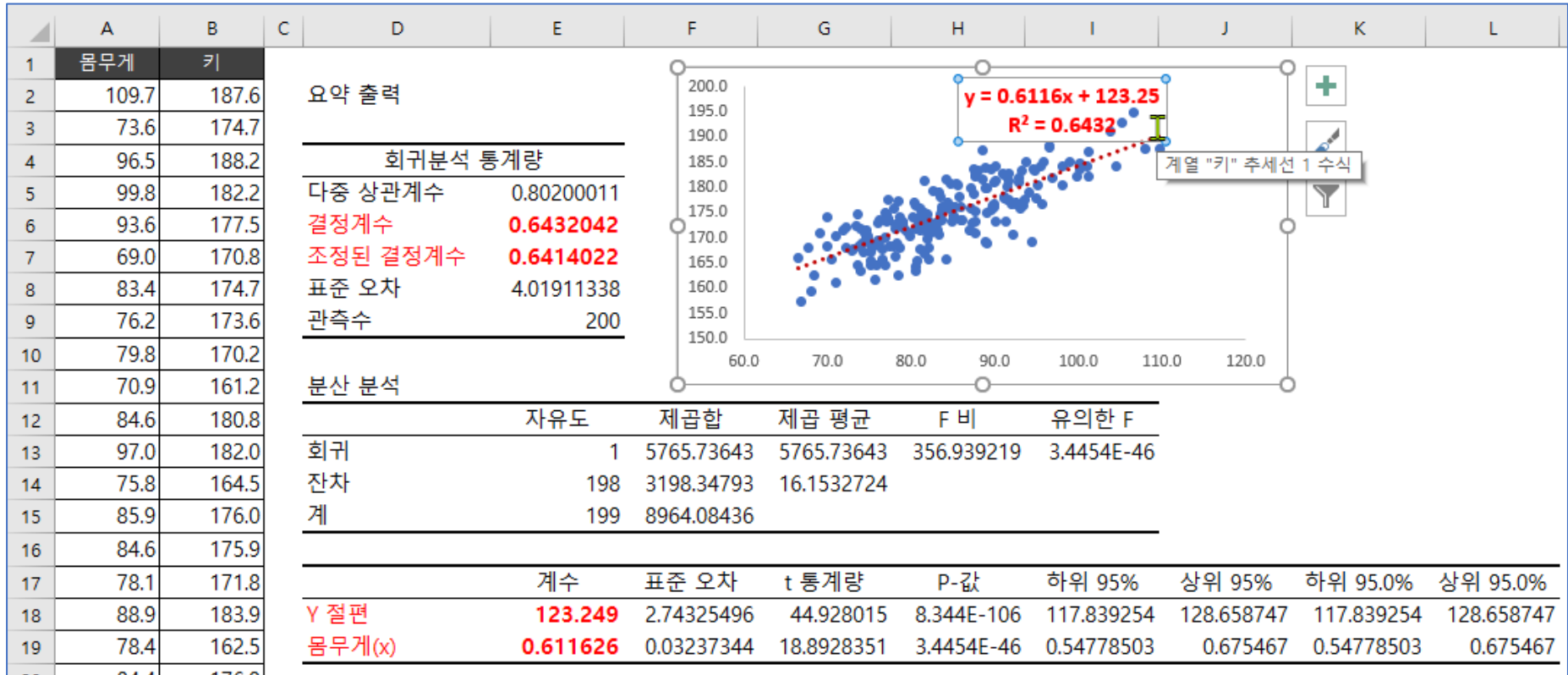
$\pm 0.81 \sim \pm 1.00$: 매우 강한 관계
 $\pm 0.61 \sim \pm 0.80$: 강한 관계
 $\pm 0.41 \sim \pm 0.60$: 보통
 $\pm 0.21 \sim \pm 0.40$: 약한 관계
 $\pm 0.00 \sim \pm 0.20$: 관계 없음 (또는 매우 약한 관계)

- Excel의 메뉴
 - [데이터]메뉴탭 → [데이터분석]메뉴
 - 분산형차트의 추세선과 수식표시
 - 상관관계 함수: CORREL(), LINEST()등



분산형 차트와 추세선 그리고 회귀식의 표시

● [데이터분석]메뉴의 [회귀분석] 결과와 분산형 차트의 추세선 식과 결정계수





학습활동: Excel 실습

실습할 Excel 파일 열기 → [데이터분석의이해실습_2.xlsx]

