Chapter 3
Descriptive Statistics II:
Numerical Descriptive Techniques

경영대학 재무금융학과
윤선중

Objectives (1)

- 기술통계량 (Descriptive Statistic)
  - 그래프 기법 (Graphical Technique)
  - 수치 방법 (Numerical Technique)
- 중심위치(central location)
  - 평균(mean), 중앙값(median), 최빈값 (mode)
- 변동성 (Variability)
  - 범위(range), 분산(variance), 표준편차(standard deviation), 변동계수(variance coefficient)
- 경험법칙 (empirical rule)
  - 평균으로부터 1표준편차, 2표준편차 의 대략적인 비율
  - 체비세프의 정리 (Chebysheff's Theorem)

Objectives (2)

- 상대위치의 척도 (Measures of Relative Standing)
  - 백분위수 (Percentile)
  - 사분위수 (Quartile)
- 변수간의 선형관계
  - 공분산(Covariance)
  - 상관계수 (Coefficient of Correlation)
  - 결정계수 (Coefficient of Determination)
  - 최소자승선 (Least Square Line)

I. Central Tendency

Terminology

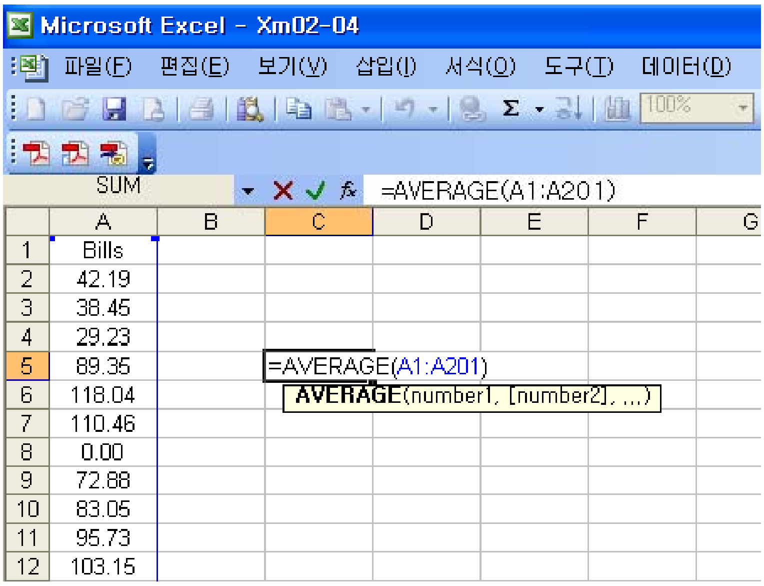
- N : 모집단의 수
- n : 표본샘플의 수
- μ : 모집단의 평균
- $\bar{x}$: 표본샘플의 평균

Mean (Arithmetic Average)

- 정의
  - 관측치들을 모두 더한 후, 이를 관측치의 개수로 나눈 값
  - 모평균: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ 표본평균: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- 산술평균 vs. 기하평균 (geometric average)
  - 기하평균: $R_g = \sqrt[n]{(1+R_1)(1+R_2)\ldots(1+R_n)} - 1$
  - 10,000원을 2년간 투자한다고 가정
  - 첫해의 수익률 100%; 이듬해의 수익률 -50%
  - 최종 투자가치는?
  - 평균수익률?
    - 산술평균과 기하평균

Mean (Arithmetic Average)

■ 예제 2.4: 평균 장거리 전화비용; Xm02-04



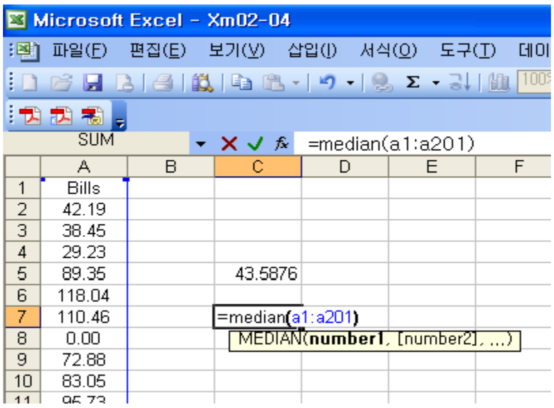
Median

■ 정의

- 모든 관측치들을 오름차순 혹은 내림차순으로 정리하였을 때, 중앙에 해당되는 관측치
- 자료의 개수가 짝수인 경우에는 중앙에 오는 두 개의 값을 더한 후, 이를 2로 나누어서 사용

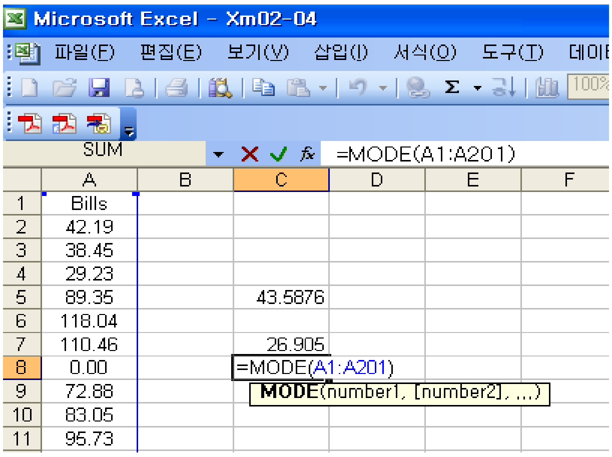
■ 예제 Xm02-04

-



Mode

- 정의
  - 관측치들 중에서 가장 큰 빈도수를 가지는 관측치
  - 최빈값은 하나의 값이 아닐 수 있음
- 예제 (Xm02-04)



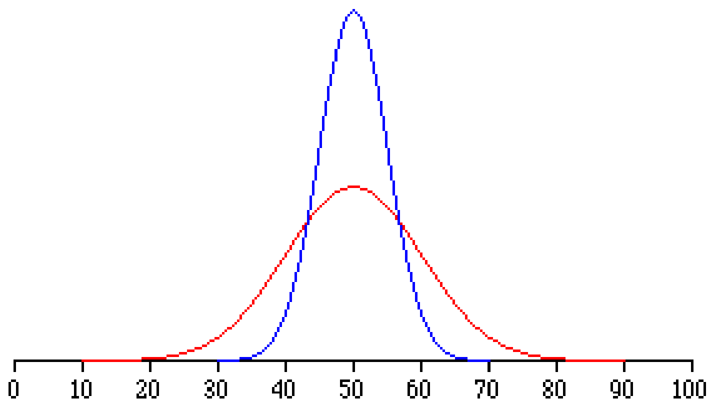
Which is better?

- 예제 1
  - 5aud의 상인들의 월 평균 수익은 다음과 같다
    - 300만원, 200만원, 100만원, 200만원, 4000만원
  - 평균 vs 평균값 vs 최빈값?
- 예제 2
  - 5명 학생의 수학과 영어 성적은 다음과 같다
    - 수학: 100, 80, 70, 50, 30
    - 영어: 75, 70, 68, 45, 25
  - 학생들이 어느 과목을 잘 하는가?
- Mean vs. Median
  - 극단치 (outlier)가 포함된 경우 Mean은 좋은 중심값이 될 수 없음
  - 계산 및 비교 등의 용도에서는 Mean이 더 좋은 특성을 가지고 있음

II. Dispersion

Measure of Dispersion

- 정의
  - 중심으로부터 흩어진 정도를 나타내는 척도
  - 아래의 두 그래프가 평균은 같음에도 불구하고 다른 산포도를 가질 수 있음

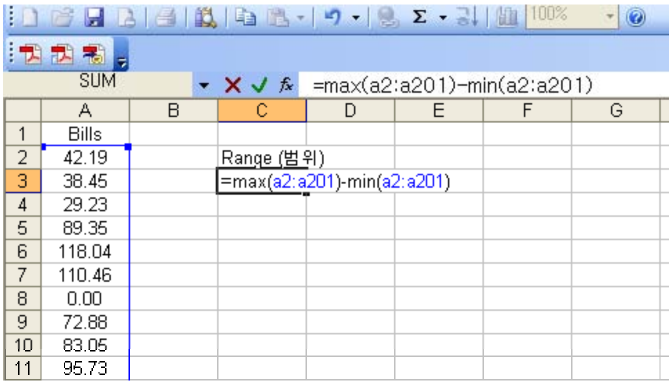


Range

■ 정의

- 관측치 중 가장 큰 값과 가장 작은 값의 차이
- 최대 관측치 – 최소 관측치

■ 예제 Xm02-04



	A	B	C	D	E	F	G
1	Bills						
2	42.19		Range (범위)				
3	38.45		=max(a2:a201)-min(a2:a201)				
4	29.23						
5	89.35						
6	118.04						
7	110.46						
8	0.00						
9	72.88						
10	83.05						
11	95.73						

Variance & Standard Deviation

■ 정의

- 편차(관측치 – 평균)를 제곱하여 더한 후 이를 관측치의 개수로 나눈 값
- 관측치들이 평균적으로 평균값에서 얼마나 떨어져 있는지 알아냄

모분산: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$ 　　표본분산: $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

모표준편차: $\sigma = \sqrt{\sigma^2}$ 　　표본표준편차: $s = \sqrt{s^2}$

■ 표준편차 vs. 변동계수 (coefficient of variation)

- 3채의 아파트 가격이 각각 1억원, 2억원, 3억원
- 3개의 주식 가격이 각각 1만원, 10만원, 100만원
- 모 변동 계수 = CV= $\dfrac{\sigma}{\mu}$ 　　표본변동계수 = cv= $\dfrac{s}{\bar{x}}$

Variance & Standard Deviation

■ 예제 3.7: 여름방학 아르바이트

- **Sample: 17, 15, 23, 7, 9, 13.**

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{17+15+23+7+9+13}{6} = \frac{84}{6} = 14\ jobs$$

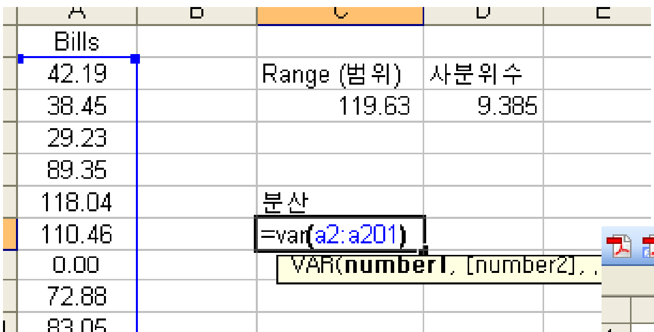
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{6-1}\left[(17-14)^2 + (15-14)^2 + \ldots(13-14)^2\right] = 33.2$$

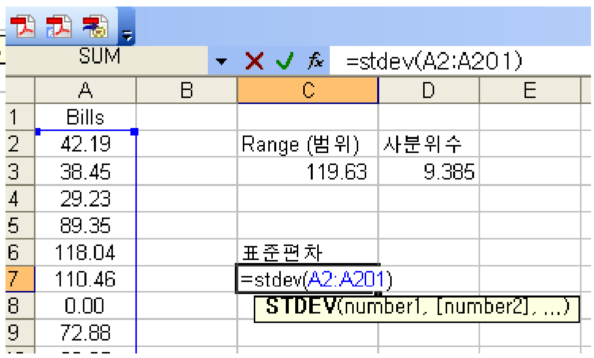
$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right] = \frac{1}{6-1}\left[\left(17^2 + 15^2 + \ldots + 13^2\right) - \frac{(17+15+\ldots+13)^2}{6}\right] = 33.2$$

Variance & Standard Deviation

■ 예제 Xm02-04

- 분산과 표준편차





Variance & Standard Deviation

- 분산과 표준편차의 특징
  - 분산과 표준편차는 항상 ___ 보다 크다
  - 모든 데이터가 같지 않는 한, 분산과 표준편차는 _____이 아님
- 경험법칙: 종모양에 국한
  - 모든 관측치의 약 68%는 평균의 1 표준편차 이내에 속한다
  - 모든 관측치의 약 95%는 평균의 2 표준편차 이내에 속한다
  - 모든 관측치의 약 99.7%는 평균의 3표준편차 이내에 속한다
  - 모든 관측치의 약 99.9996%는 평균의 6표준편차 이내에 속한다

Variance & Standard Deviation

- 체비세브의 정리 (Chebysheff's Theorem)
  - 모든 히스토그램의 모습에 적용 (경험의 법칙보다 더 일반적임)

$$1-\frac{1}{k^2} \quad for \quad k>1$$

- 측정 단위 혹은 규모가 다른 두 자료의 비교
  - 표준화의 활용 (예: 표준점수)

$$\boxed{Z=\frac{X-\mu}{\sigma}}$$

  - 변동계수의 활용

Relative Location

- 백분위수 (Pth percentile)
  - 이 값보다 적은 값들이 관측치들의 P%이고 이 값보다 큰 값들이 관측치들의 (100-P)%인 값
- 백분위수의 위치

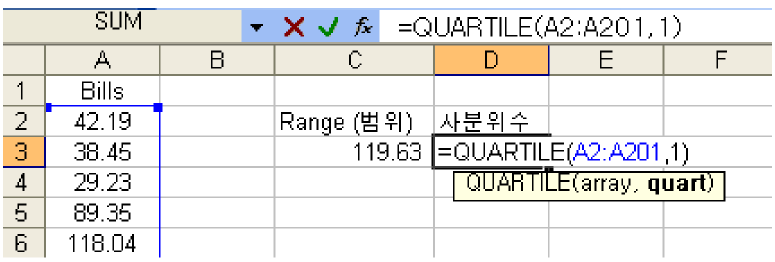
$$L_P = (n+1)\frac{P}{100}$$

$$\text{where } L_P \text{ is the location of the } P^{th} \text{ percentile}$$

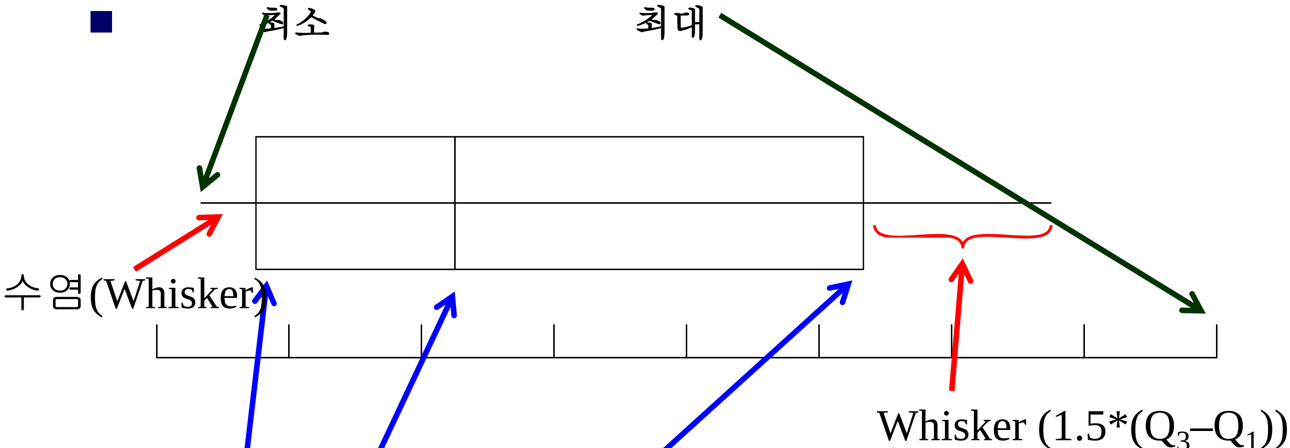
- 예제 3.11
  - Sample: 0 0 5 7 8 9 12 14 22 33
  - L_{25} = (10+1)(25/100) = 2.75: 2번째와 3번째 수의 ¾ 위치에 존재
  - 0 + 3.75 = **3.75**

Quartiles

- 정의
  - 사분위수
  - 데이터를 같은 크기의 네부분으로 나누어 흩어진 정보를 측정
  - 첫번째 사분위수 (first quartiles): 상위 25%
  - 두번째 사분위수 (second quartiles): 상위 50%
  - 참고: 십분위수(deciles), 백분위수 (percentiles)
  - 사분위 범위 (inter quartile range): Q3 - Q1
- 예제



Box Plot

- 박스그림(Box plot)은 5개의 통계치 (최소, 최대, 첫사분위수, 두번째 사분위수, 세번째사분위수)를 그림에 담을 수 있다.
- 

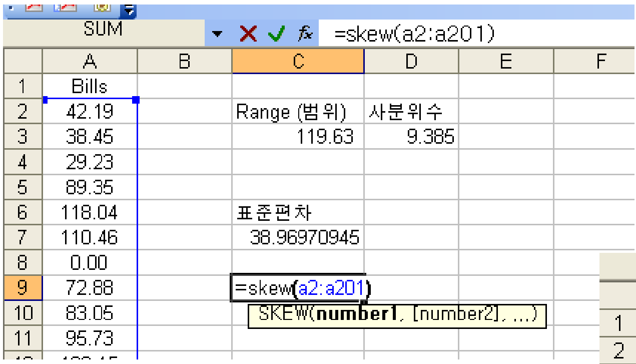
  - 첫번째, 두번째, 세번째 사분위 수
- Data Analysis Plus를 통해 구현 가능

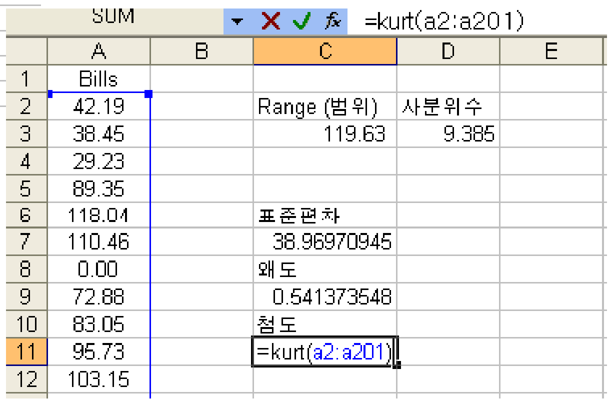
Skewness & Kurtosis

- 왜도 (skewness)
  - 평균을 중심으로 어느 정도 대칭적인 모양인가를 나타내는 척도
  - (1) 왜도<0: ___________으로 긴 꼬리
  - (2) 왜도>0: ___________으로 긴 꼬리
  - (3) 왜도=0: __________임을 의미
- 첨도 (kurtosis)
  - 데이터들의 분포가 정규분포에 비해 얼마나 뾰족한지를 나타내는척도
  - (1) 첨도<3: 정규분포보다 _____한 분포
  - (2) 첨도>3: 정규분포보다 _____한 분포
  - (3) 첨도=3: 정규분포보다 _____한 분포

Skewness & Kurtosis

■ 예제 Xm02-04

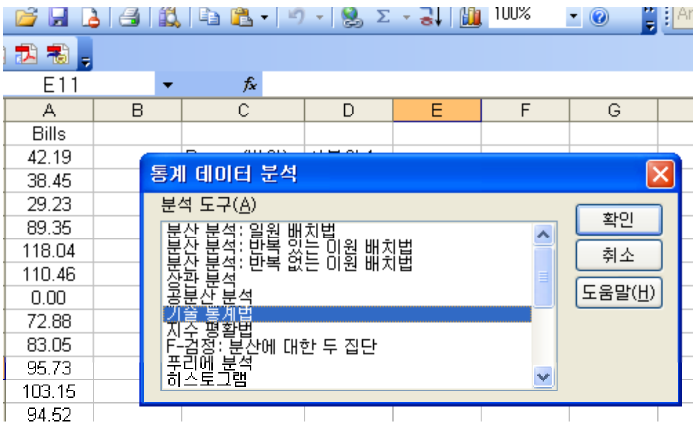




Summary Statistics

■ 예제 Xm02-04

- 도구 – 데이터분석-기술통계법
- 입력범위 A1:A201
- 요약통계량 클릭



Column1	
평균	43.5876
표준 오차	2.755575
중앙값	26.905
최빈값	0
표준 편차	38.96971
분산	1518.638
첨도	-1.291907
왜도	0.541374
범위	119.63
최소값	0
최대값	119.63
합	8717.52
관측수	200

III. Linear Relationship

- 두 변수간의 선형관계의 강도와 방향(strength & direction) 에 대한 정보를 제공하는 수치 기법에 대해 살펴봄
  - 공분산 (covariance),
  - 상관계수 (coefficient of correlation)
  - 결정계수 (coefficient of determination)

Covariance

- 정의
  - 두 변수의 값이 각각의 평균으로부터 얼마나 떨어져 있는지를 나타내는 수치로써, 두 변수간의 선형관계를 파악하기 위해서 사용

$$\text{Population covariance} = \sigma_{xy} = \frac{\sum_{i=1}^{N}(\mathrm{x}_i - \mu_x)(y_i - \mu_y)}{\mathrm{N}}$$

$$\text{Sample covariance} = \mathrm{s}_{xy} = \frac{\sum_{i=1}^{n}(\mathrm{x}_i - \bar{x})(y_i - \bar{y})}{\mathrm{n - 1}} \qquad s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

- 공분산의 특징
  - (-inf, inf) 사이의 값을 가짐
  - 공분산은 ____에 따라 두 변수간의 상관방향만을 나타낼 뿐이고, 크기는 무관

Correlation Coefficient

- 정의
  - 공분산을 두 변수의 표준편차의 곱으로 나누어준 값
  - 공분산에는 두 변수의 scale에 대한 고려가 전혀 없는데 반해, 상관계수는 scale에 대한 고려가 포함되어 있음

Population coefficient of correlation: $\rho = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$ Sample coefficient of correlation: $r = \dfrac{s_{xy}}{s_x s_y}$

- 특징
  - 상관계수는 항상 ___과 ___사이에 존재함
  - 상관계수가 1에 가까우면, ______ 상관관계를 의미
  - 상관계수가 -1에 가까우면, ______상관관계를 의미
  - 상관계수가 0에 가까우면, ______상관관계 의미
  - (1) 두 변수가 독립이면, 두 변수간의 상관계수는 0?
  - (2) 두 변수간의 상관계수가 0이면, 두 변수는 독립?

Example

- 예제 3.16: 공분산의 계산

	X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	covariance
	2	13	-3	-7	21	
Set #1	6	20	1	0	0	S_{xy} = 17.5
	7	27	2	7	14	
	2	27	-3	7	-21	
Set #2	6	20	1	0	0	S_{xy} = -17.5
	7	13	2	-7	-14	
	2	20	-3	0	0	
Set #3	6	27	1	7	7	S_{xy} = -3.5
	7	13	2	-7	-14	

For each set: $\bar{X} = 5$ $\bar{Y} = 20$

Example

■ 표준편차의 계산

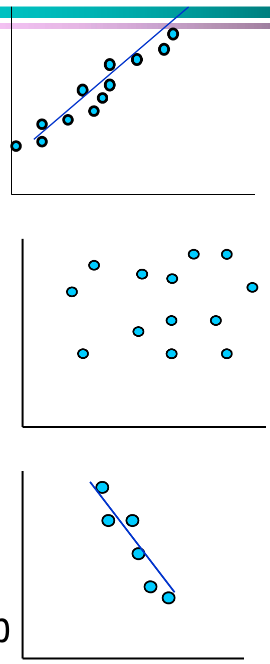
$$r = \frac{s_{xy}}{s_x s_y} = \frac{17.5}{(2.65)(7.0)} = .943$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-17.5}{(2.65)(7.0)} = -.943$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-3.5}{(2.65)(7.0)} = -.189$$

Correlation Coefficient vs. Scatter Diagram

$$\rho \text{ or } r = \begin{cases} +1 & \text{Strong positive linear relationship} \\ 0 & \text{No linear relationship} \\ -1 & \text{Strong negative linear relationship} \end{cases}$$



IV. Least Square Method

- 산포도는 선형관계의 강도와 방향을 측정
- 산포도에 직선을 그어 그 강도와 방향을 추정
- 이를 위해 개발된 방법이 최소자승법(Least Square Method)

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$\hat{y} = b_0 + b_1 x$$

Least Squares Method

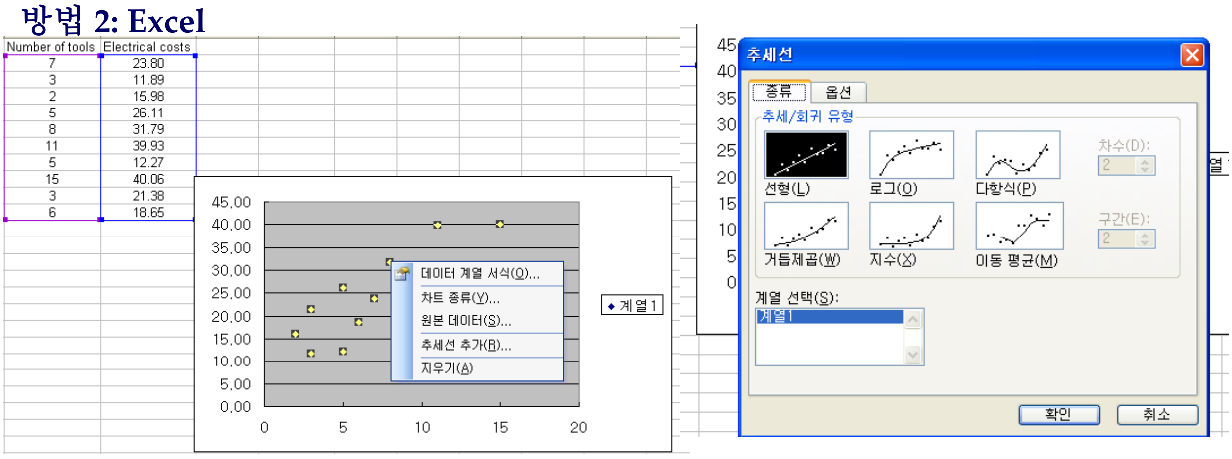
- 예제 3.17: 고정비용과 변동비용의 추정; Xm03-17

$$y = b_0 + b_1 x$$

- 단, y = 전체비용, b_0 = 고정비용, b_1 =변동비용, x =공구수

방법 1: 직접 계산: 교제 125페이지 참조

방법 2: Excel



Least Squares Method