

3 장, Part A

기술통계: 수치 척도

- 위치 척도
- 변동성 척도

위치척도(measures of location)

- 평균
- 중앙값
- 최빈값
- 백분위수
- 사분위수

척도들이 표본으로부터 추출된 자료로부터 계산된다면, 이를 표본통계량(sample statistic)이라 한다.

척도들이 모집단의 자료로부터 도출된다면, 모집단 모수(population parameter)라고 한다.

표본통계량은 해당 모수의 점 추정량(point estimator)이다.

평균(mean)

- 자료집합의 평균은 모든 자료값들의 평균이다.
- 표본평균 \bar{x} 은 모집단 평균 μ 의 점추정량이다.

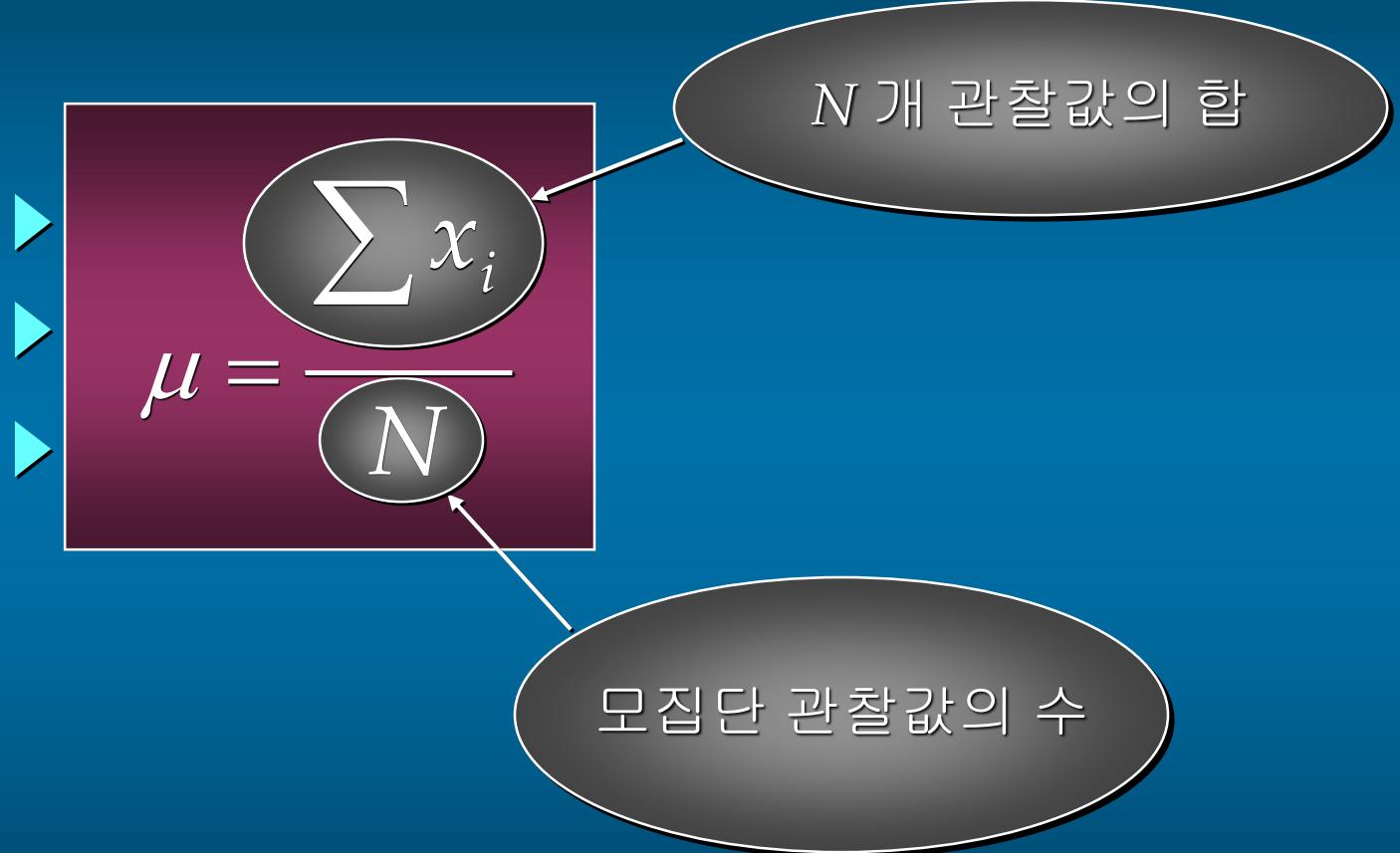
표본평균(sample mean) \bar{x}

$$\bar{x} = \frac{\sum x_i}{n}$$

n개의 관찰값의 합

표본 관찰값의 수

모집단 평균(population mean) μ



표본 평균

■ 예 : 아파트 임대

어느 대학가에서 표본으로 간이
아파트 (efficiency apartments)
70채가 무작위 선정되었다.
이 아파트의 월세는
오름차순으로
다음 슬라이더에 나타나 있다.



표본 평균



425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

표본 평균



$$\bar{x} = \frac{\sum x_i}{n} = \frac{34,356}{70} = 490.80$$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

중앙값(median)

- ▶ 중앙값은 자료가 순서대로(오름차순) 배열되어 있을 때 중앙에 있는 값을 의미한다.
- ▶ 자료에 극단값이 포함되어 있을 경우, 중앙값은 중심위치를 측정하는 데에 있어서 선호된다
- ▶ 연소득이나 재산 자료에서는 중앙값이 위치척도로 자주 사용된다.
- ▶ 몇몇의 극단값을 갖는 소득이나 재산은 평균을 부풀릴 수 있다.

중앙값

- 자료의 수가 홀수일 경우:

26	18	27	12	14	27	19
----	----	----	----	----	----	----

7개 관찰값

12	14	18	19	26	27	27
----	----	----	----	----	----	----

오름차순

중앙값은 가운데 값이다.

중앙값 = 19

중앙값

- 자료의 수가 짝수일 경우:

26	18	27	12	14	27	30	19
----	----	----	----	----	----	----	----

8 개 관찰값

12	14	18	19	26	27	27	30
----	----	----	----	----	----	----	----

오름차순

중앙값은 중앙에 있는 두 개 숫자의 평균이다.

$$\text{중앙값} = (19 + 26)/2 = 22.5$$

중앙값



35, 36번째 값의 평균:

$$\text{중앙값} = (475 + 475)/2 = 475$$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

최빈값(mode)

- ▶ 최빈값은 가장 빈번하게 나타나는 값이다.
- ▶ 가장 빈번하게 나타나는 값이 2개 이상이 있을 수도 있다.
- ▶ 만약 자료가 2개의 최빈값을 가진다면, 이를 이중모드(bimodal)라고 한다.
- ▶ 만약 자료가 3개 이상의 최빈값을 가진다면, 이를 다모드(multimodal)라고 한다.

최빈값



450 이 가장 많다 (7 번)

최빈값 = 450

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

백분위수(percentiles)

- ▶ 백분위수는 가장 작은 값부터 가장 큰 값 사이에 자료가 어떻게 퍼져 있는지에 대한 정보를 제공한다.
- ▶ 대학의 입학시험성적은 주로 백분위수로 보고된다.

백분위수

- p 백분위수는 적어도 관찰값의 p 퍼센트가 그 값과 같거나 작은 값이다. 그리고 적어도 관찰값의 $(100-p)$ 퍼센트는 p 백분위수와 같거나 더 큰 값을 가진다.

백분위수

▶ 자료를 크기 순서대로 배열한다(오름차순).

▶ 지표 (index) i , p 백분위수의 위치를 계산한다.

$$i = (p/100)n$$

▶ 만약 i 가 정수가 아니라면, 올림한다. p 백분위수는 i 번째 위치한 값이다.

▶ 만약 i 정수라면, p 백분위수는 i 와 $i+1$ 번째 위치한 값의 평균이다.

80th 백분위수



$$i = (p/100)n = (80/100)70 = 56$$

56, 57번째 값의 평균:

$$80 \text{ 백분위수} = (535 + 549)/2 = 542$$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

80th 백분위수



▶ “적어도 관찰값의
80%가
542 보다 작거나
같다.”

▶ “적어도 관찰값의
20%가
542 보다 크거나
같다.”

▶ $56/70 = .8$ 또는 80%

▶ $14/70 = .2$ 또는 20%

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

사분위수(quartiles)

- ▶ 사분위수는 특정 백분위수이다.
- ▶ 1사분위수 = 25th 백분위수
- ▶ 2사분위수 = 50th 백분위수 = 중앙값
- ▶ 3사분위수 = 75th 백분위수

3사분위수



3사분위수 = 75th 백분위수

$$i = (p/100)n = (75/100)70 = 52.5 = 53$$

3사분위수 = 525

425	430	430	435	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465	
465	470	470	472	475	475	475	480	480	480	
480	485	490	490	490	500	500	500	500	510	
510	515	525	525	525	535	549	550	570	570	
575	575	580	590	600	600	600	600	615	615	

변동성 측정(measures of variability)

- ▶ 위치를 측정하는 것 외에도 변동성(산포도)을 측정하는 것이 필요할 때가 있다.
- ▶ 예를 들어, 공급자 A 또는 공급자 B를 선정 할 때, 각각의 평균 배달기간 뿐만 아니라 각각의 배달기간 변동성도 고려하여야 한다.

변동성 측정

- ▶ 범위
- ▶ 사분위수 범위
- ▶ 분산
- ▶ 표준편차
- ▶ 변동계수

범위(range)

- ▶ 범위는 최대값과 최소값의 차이이다.
- ▶ 변동성을 측정하는 가장 단순한 방법이다.
- ▶ 범위는 최대값과 최소값에 영향을 많이 받는다.

범위



범위 = 최대값 - 최소값

$$= 615 - 425 = 190$$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

사분위수 범위(interquartile range)

- ▶ 사분위수 범위는 3사분위수와 1사분위수의 차이이다.
- ▶ 사분위수 범위는 자료의 중앙 50%의 범위를 의미한다.
- ▶ 이는 극단값의 영향을 줄일 수 있다.

사분위수 범위



3사분위수 ($Q3$) = 525

1사분위수 ($Q1$) = 445

사분위수 범위 = $Q3 - Q1 = 525 - 445 = 80$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

분산(variance)

- ▶ 분산은 자료 모두를 사용하여 그 자료의 변동성을 측정하는 도구이다.
- ▶ 분산은 각각의 관찰값 (x_i)과 평균 (표본평균 \bar{x} , 모집단평균 μ) 과의 차이에 기초한다.

분산

- ▶ 분산은 각각의 자료값과 평균과의 차이에 대한 제곱의 평균이다.
- ▶ 분산은 아래와 같이 구한다:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

표본 분산

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

모집단 분산

표준편차(standard deviation)

- ▶ 표준편자는 분산에 제곱근(square root)을 취한 값이다.
- ▶ 표준편자는 원래의 자료에서 사용된 단위와 동일한 단위로 측정되므로 분산보다 해석이 용이하다.

표준편차

표준편자는 아래와 같이 계산된다:

$$s = \sqrt{s^2}$$

$$\sigma = \sqrt{\sigma^2}$$

표본표준편자

모집단표준편자

변동계수(coefficient of variance)

변동계수는 표준편차가 평균에 비하여 얼마나 큰지를 나타낸다.

변동계수는 아래와 같이 계산된다:

$$\left(\frac{s}{\bar{x}} \times 100 \right) \%$$

$$\left(\frac{\sigma}{\mu} \times 100 \right) \%$$

표본의 경우

모집단의 경우

분산, 표준편차, 변동계수



▶ 분산

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 2,996.16$$

▶ 표준편차

$$s = \sqrt{s^2} = \sqrt{2996.47} = 54.74$$

표준편자는
평균의 약
11% 이다.

▶ 변동계수

$$\left(\frac{s}{\bar{x}} \times 100 \right) \% = \left(\frac{54.74}{490.80} \times 100 \right) \% = 11.15\%$$

3 장, Part B

기술 통계: 수치 척도

- ▶ 분포 형태, 상대적 위치, 극단값
- ▶ 탐색적 자료분석
- ▶ 두 변수간의 관련성 측정
- ▶ 가중평균과 그룹화 자료

분포 형태, 상대적 위치, 극단값

- 분포 형태
- z-값
- 체비셰프의 원리
- 경험법칙
- 극단값 찾기

분포 형태: 왜도(skewness)

- ▶ 분포 형태를 측정하는 중요한 척도 중 하나를 ‘왜도’라고 한다.
- ▶ 자료집합의 왜도를 구하는 계산식은 조금 복잡하다.

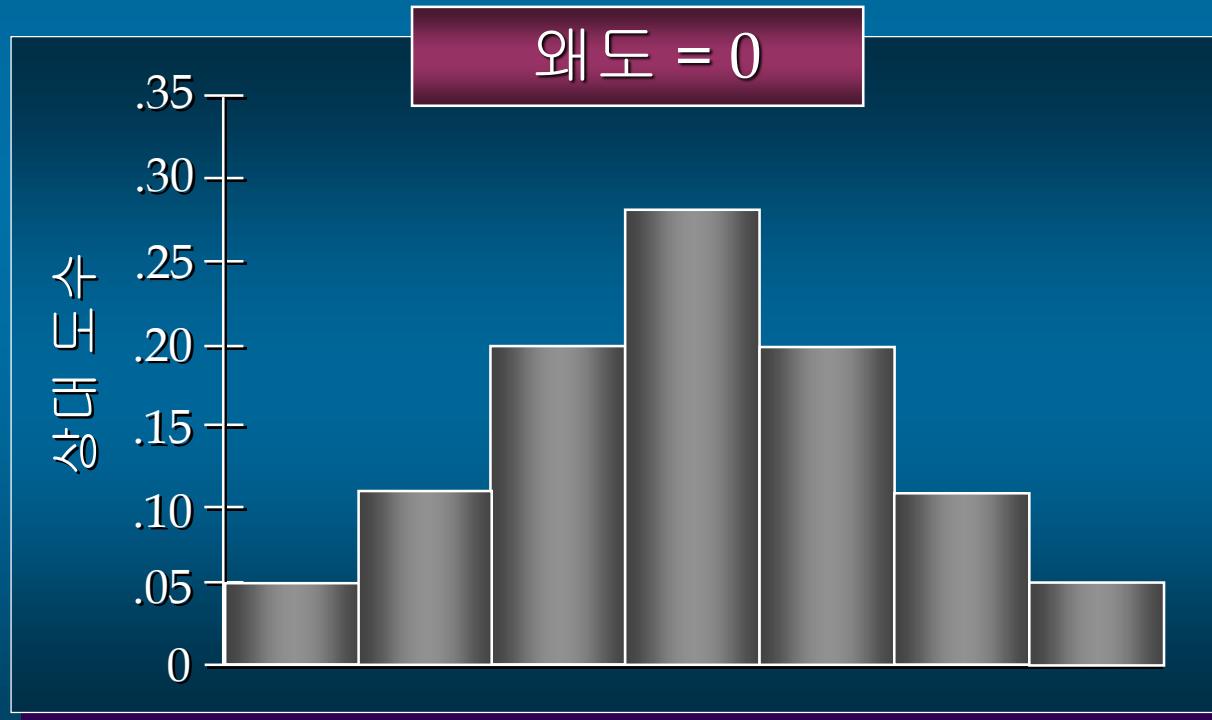
$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- ▶ 통계프로그램을 사용하여 왜도를 쉽게 계산할 수 있다.

분포 형태: 왜도

■ 정대칭(한 쪽으로 치우치지 않음)

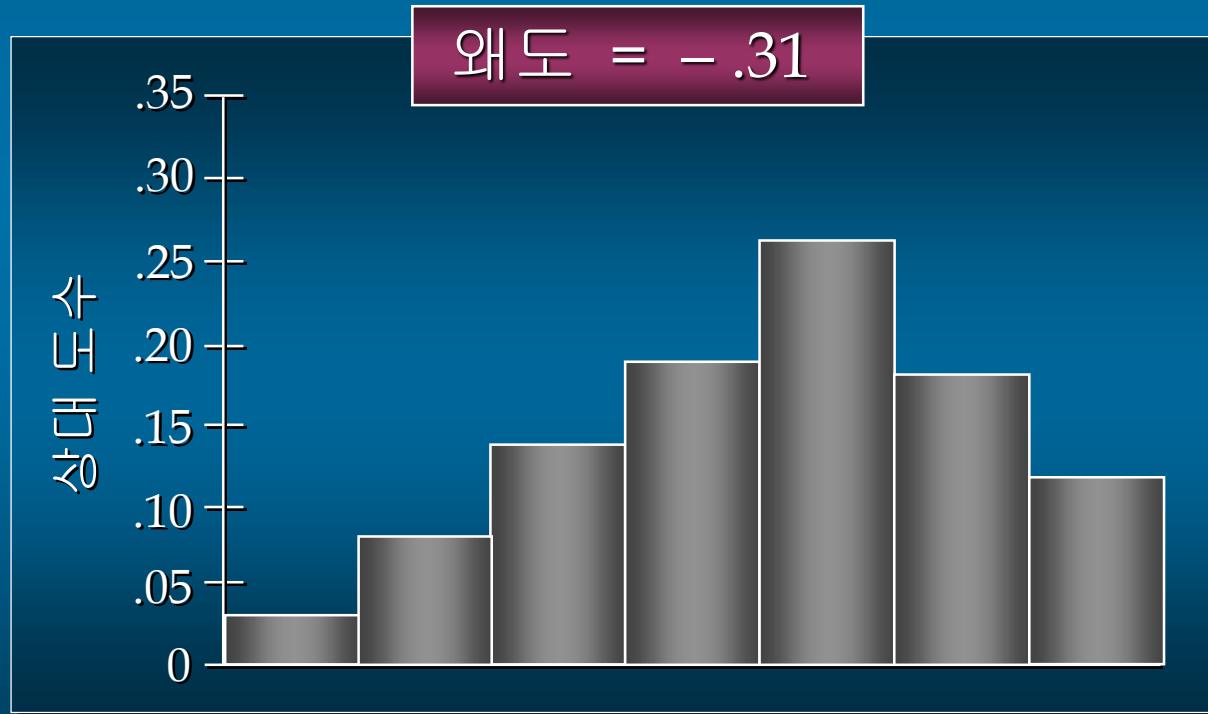
- 왜도는 0이다.
- 평균과 중앙값은 같다.



분포 형태: 왜도

■ 왼쪽으로 치우친 경우(왼쪽 꼬리분포)

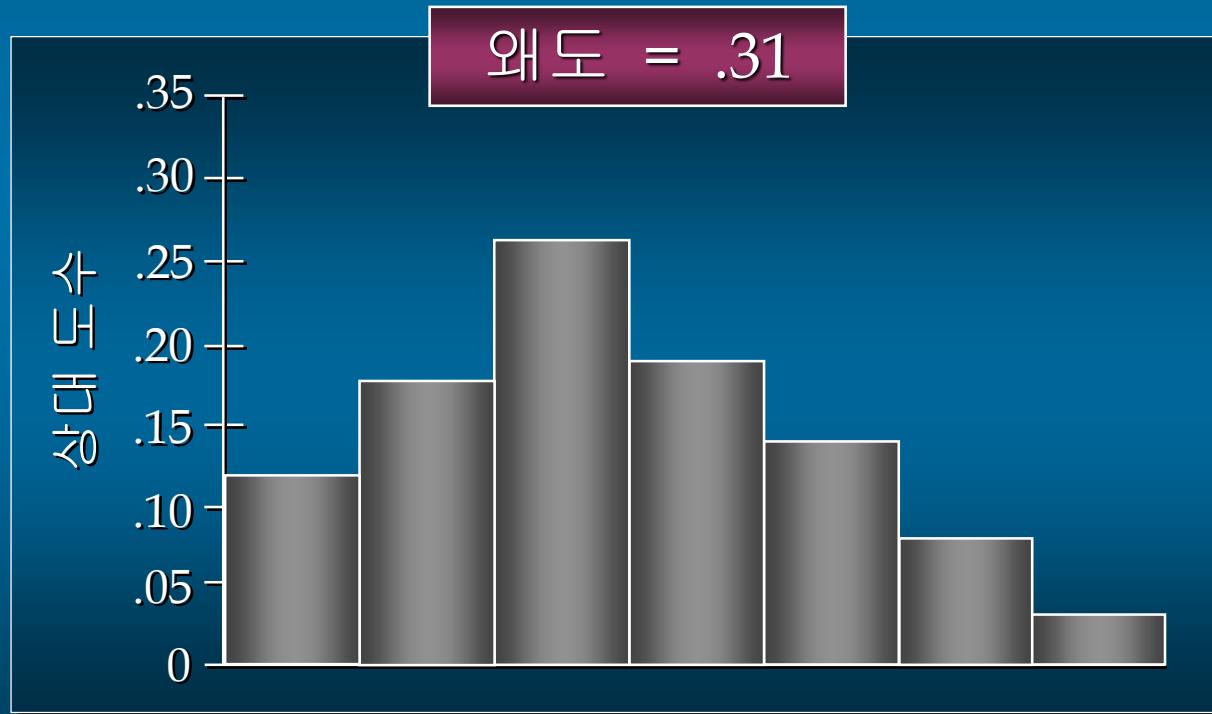
- 왜도는 음(-)이다.
- 보통 평균은 중앙값 보다 작다.



분포 형태: 왜도

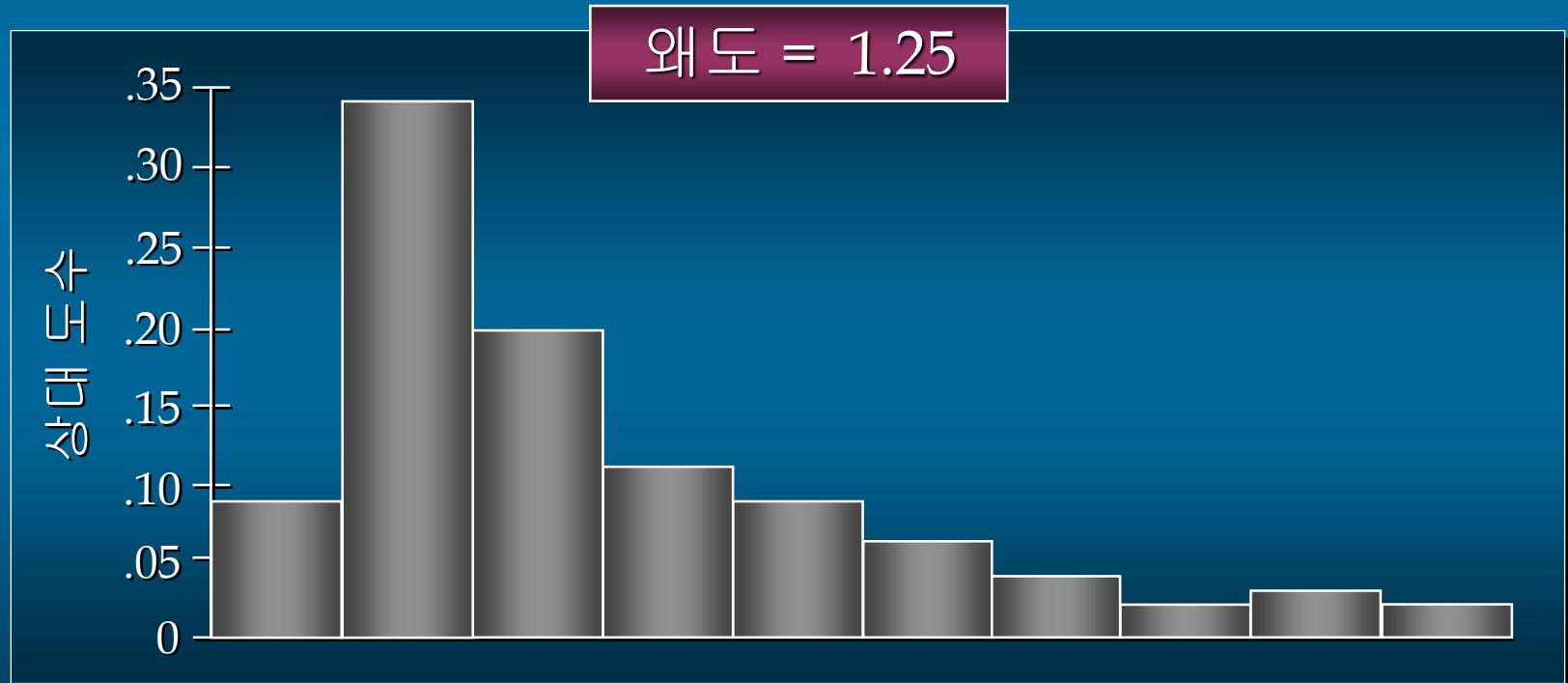
■ 오른쪽으로 치우친 경우(오른쪽 꼬리분포)

- 왜도는 양(+)이다.
- 보통 평균은 중앙값 보다 크다.



분포 형태: 왜도

- 오른쪽으로 심하게 치우친 경우(심한 오른쪽 고리분포)
 - 왜도는 양(+)이다 (종종 1.0보다 높다).
 - 보통 평균은 중앙값 보다 크다.



Z-값

- ▶ Z-값을 종종 ‘표준화(된) 값’이라고 한다.
- ▶ 이는 관찰값 x_i 와 평균과의 거리가 표준편차의 몇 배에 해당하는지를 나타낸다.

$$z_i = \frac{x_i - \bar{x}}{s}$$

z -값

- ▶ 관찰값의 z -값은 자료에서 해당 관찰값의 상대 위치를 측정하는 척도이다.
- ▶ 표본평균보다 작은 자료값은 0보다 작은 z -값을 갖는다.
- ▶ 표본평균보다 큰 자료값은 0보다 큰 z -값을 갖는다.
- ▶ 표본평균과 같은 자료값은 0인 z -값을 갖는다.

z-값



■ 가장 작은 값(425)의 z-값

$$z = \frac{x_i - \bar{x}}{s} = \frac{425 - 490.80}{54.74} = -1.20$$

월세값의 표준화 값										
-1.20	-1.11	-1.11	-1.02	-1.02	-1.02	-1.02	-1.02	-0.93	-0.93	-0.93
-0.93	-0.93	-0.93	-0.84	-0.84	-0.84	-0.84	-0.84	-0.75	-0.75	-0.75
-0.75	-0.75	-0.75	-0.75	-0.75	-0.56	-0.56	-0.56	-0.47	-0.47	-0.47
-0.47	-0.38	-0.38	-0.34	-0.29	-0.29	-0.29	-0.20	-0.20	-0.20	-0.20
-0.20	-0.11	-0.01	-0.01	-0.01	0.17	0.17	0.17	0.17	0.35	
0.35	0.44	0.62	0.62	0.62	0.81	1.06	1.08	1.45	1.45	
1.54	1.54	1.63	1.81	1.99	1.99	1.99	1.99	2.27	2.27	

체비셰프의 정리



어떤 자료에 있는 항목들의 적어도 $(1 - 1/z^2)$ 의 값은 평균에서 z 표준 편차 크기의 범위 안에 있어야 한다.
그리고 이 때의 z는 1보다 더 큰 값이다.

$$P\{|X-\mu| \leq z\sigma\} \geq 1 - 1/z^2$$

체비셰프의 원리(Chebyshev's theorem)

- ▶ 적어도 자료값들의 **75%** 는 평균에서
 $z = 2$ 표준편차 범위 안에 있어야 한다.

- ▶ 적어도 자료값들의 **89%** 는 평균에서
 $z = 3$ 표준편차 범위 안에 있어야 한다.

- ▶ 적어도 자료값들의 **94%** 는 평균에서
 $z = 4$ 표준편차 범위 안에 있어야 한다.

체비셰프의 원리



예:

$$z = 1.5 \quad (\bar{x} = 490.80 \text{ 와 } s = 54.74)$$

적어도 월세값들의

$(1 - 1/(1.5)^2) = 1 - 0.44 = 0.56$ 또는 56%
는 아래의 값들 사이에 있어야 한다.

$$\bar{x} - z(s) = 490.80 - 1.5(54.74) = 409$$

와

$$\bar{x} + z(s) = 490.80 + 1.5(54.74) = 573$$

(실제, 86%의 월세값들이 409 와 573사이에 있다.)

경험 법칙(Empirical rule)

자료가 대략 종모양의 분포를 띠 것으로 생각되면,

- ▶ 경험법칙은 평균의 특정 표준편차 내에 있어야 하는 자료값들의 비율을 결정하는데 사용될 수 있다.

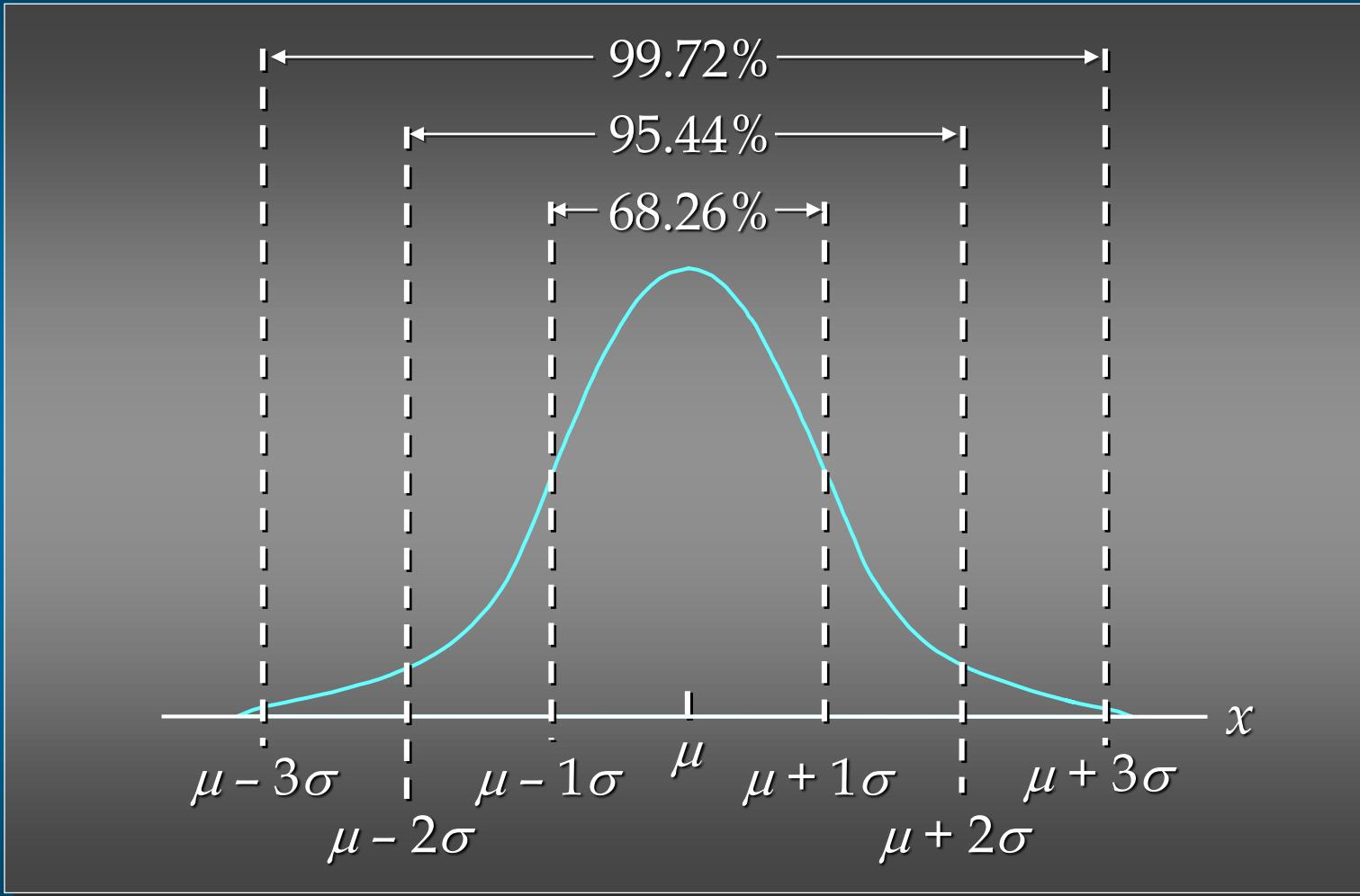
- ▶ 경험법칙은 제6장에서 다루어 질 정규분포에 기반한다.

경험 법칙

종모양 분포를 가지는 자료에 대하여:

- ▶ 정규확률변수값의 **68.26%** 가 평균의
 ± 1 표준편차 범위 안에 있다.
- ▶ 정규확률변수값의 **95.44%** 가 평균의
 ± 2 표준편차 범위 안에 있다.
- ▶ 정규확률변수값의 **99.72%** 가 평균의
 ± 3 표준편차 범위 안에 있다.

경험 법칙



극단값 찾기(detecting outliers)

- ▶ 극단값은 자료에서 특출나게 작거나 큰 값을 말한다.
- ▶ -3보다 작거나 +3보다 큰 z-값에 해당하는 자료값을 극단값으로 보면 된다.
- ▶ 극단값은 다음과 같은 경우에 생긴다.
 - 잘못 기록된 자료값
 - 자료에 잘못 포함된 값
 - 자료에 제대로 포함된 제대로 기록된 값

극단값 찾기



가장 극단적 z-값은 -1.20 과 2.27

극단점 기준으로 $|z| \geq 3$ 을 사용하면, 이 자료에는
극단값이 없다.

월세에 대한 표준화 값

-1.20	-1.11	-1.11	-1.02	-1.02	-1.02	-1.02	-1.02	-0.93	-0.93
-0.93	-0.93	-0.93	-0.84	-0.84	-0.84	-0.84	-0.84	-0.75	-0.75
-0.75	-0.75	-0.75	-0.75	-0.75	-0.56	-0.56	-0.56	-0.47	-0.47
-0.47	-0.38	-0.38	-0.34	-0.29	-0.29	-0.29	-0.20	-0.20	-0.20
-0.20	-0.11	-0.01	-0.01	-0.01	0.17	0.17	0.17	0.17	0.35
0.35	0.44	0.62	0.62	0.62	0.81	1.06	1.08	1.45	1.45
1.54	1.54	1.63	1.81	1.99	1.99	1.99	1.99	2.27	2.27

탐색적 자료분석(Exploratory data analysis)

- 다섯 수치 요약
- 상자 그림

다섯 수치 요약(Five-number summary)

1	최소값
2	1사분위수
3	중앙값
4	3사분위수
5	최대값

다섯 수치 요약



최소값 = 425

1사분위수 = 445

중앙값 = 475

3사분위수 = 525

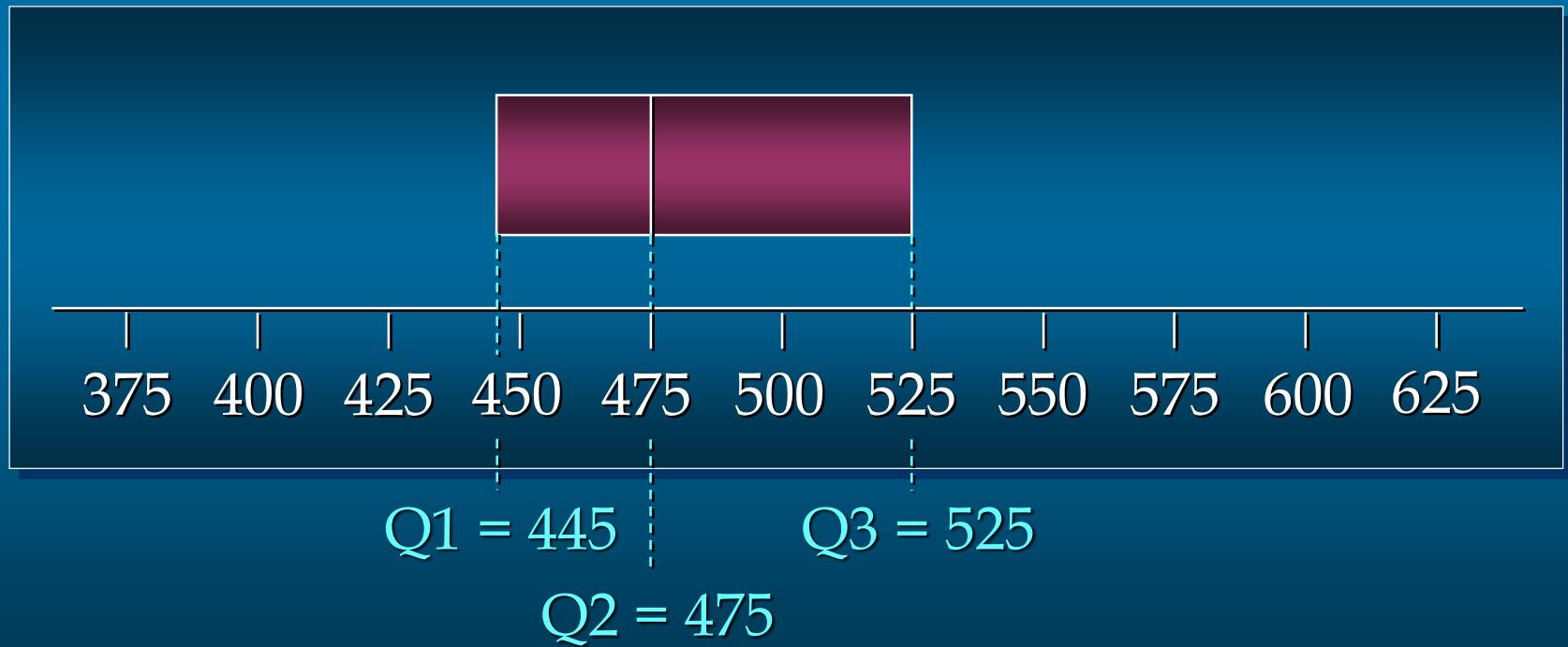
최대값 = 615

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

상자 그림(Box Plot)



- ▶ 상자의 양 끝은 1사분위수와 3사분위수에 위치하게 한다.
- ▶ 상자안에 수직선을 중앙값 위치에 그린다 (2사분위수).



상자 그림



- 사분위수간 범위($IQR=Q_3-Q_1$)를 사용하여 상한선과 하한선을 그린다.
- 이 범위 밖의 자료는 극단값이라고 할 수 있다.
- 각 극단값의 위치는 *로 표시한다.

... 계속됨 →

상자 그림



- ▶ 하한선은 $Q1$ 보다 아래쪽 $1.5(IQR)$ 이다.

$$\text{하한선: } Q1 - 1.5(IQR) = 445 - 1.5(80) = 325$$

- ▶ 상한선은 $Q3$ 보다 위쪽 $1.5(IQR)$ 이다.

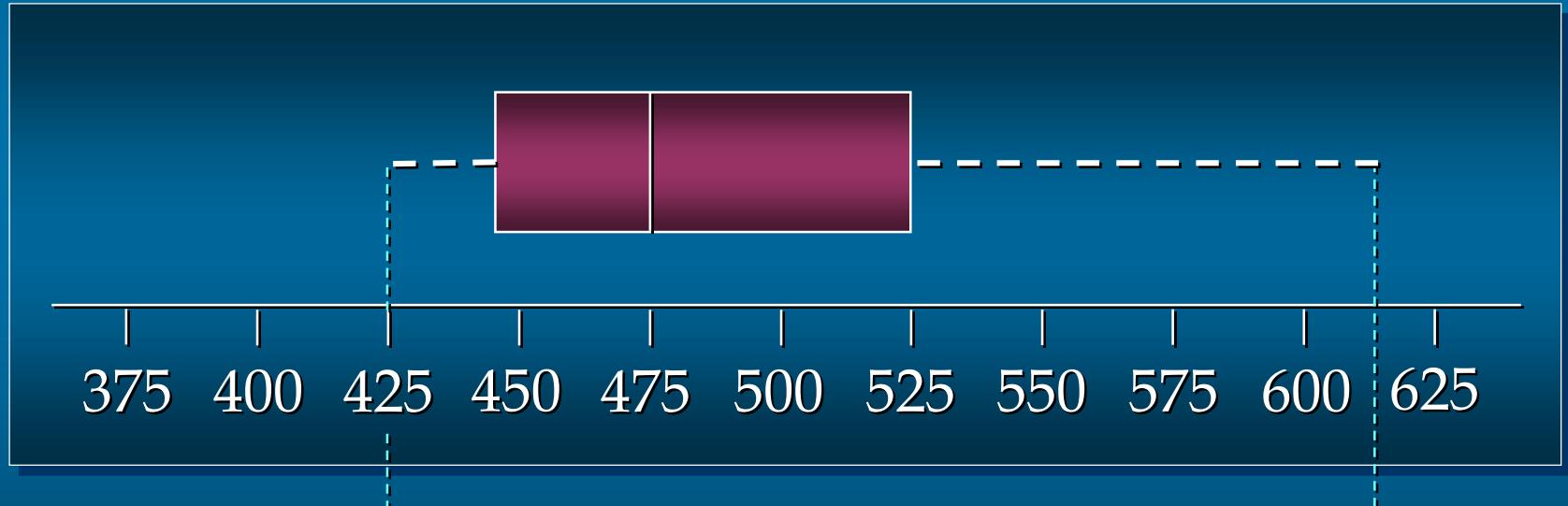
$$\text{상한선: } Q3 + 1.5(IQR) = 525 + 1.5(80) = 645$$

- ▶ 아파트 월세자료에서 극단값 (325 보다 작거나 645 보다 큰 값)은 없다.

상자 그림



- ▶ 상자의 양끝에서 한계선내에 최소값과 최대값까지 점선(whiskers)을 그린다.



한계선내 최소값 = 425

한계선내 최대값 = 615

두 변수간의 연관성 측정 (Measures of association between two variables)

- ▶ 지금까지 하나의 변수에 대한 자료를 요약하기 위한 수치적 방법들을 살펴보았다.
- ▶ 경영자나 의사결정자는 종종 두 변수의 관계에 관심을 가진다.
- ▶ 두 변수의 관계에 관한 기술 측정치로 공분산과 상관계수가 있다.

공분산(covariance)

- ▶ 공분산은 두 변수의 선형관계를 측정하는 척도이다.
- ▶ 양의 값은 양의 관계를 나타낸다.
- ▶ 음의 값은 음의 관계를 나타낸다.

공분산

공분산은 아래와 같이 계산된다:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

표본의 경우

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

모집단의 경우

상관계수(correlation coefficient)

- ▶ 상관계수는 -1에서 +1사이의 값을 갖는다.
- ▶ -1값에 가까울수록 강한 음의 선형관계를 나타낸다.
- ▶ +1값에 가까울수록 강한 양의 선형관계를 나타낸다.

상관계수

상관계수는 아래와 같이 계산된다:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

표본의 경우

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

모집단의 경우

상관계수

상관관계는 변수들 간의 선형관계를 측정하는 것이지 반드시 인과관계를 측정하는 것은 아니다.

두 변수가 높은 상관관계를 갖는다고 해도, 한 변수가 다른 변수의 원인이 된다는 것을 의미하지는 않는다.
예, 식당의 일반적인 식사가격과 음식의 질

공분산과 상관계수

예: 어떤 골프선수가 드라이빙 거리와 18홀 점수간에 서로 관계가 있는지에 대하여 조사하고자 한다.



평균 드라이빙 거리 (yds.)	평균 18홀 점수
277.6	69
259.5	71
269.1	70
267.0	70
255.6	71
272.9	69

공분산과 상관계수



x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
277.6	69	10.65	-1.0	-10.65
259.5	71	-7.45	1.0	-7.45
269.1	70	2.15	0	0
267.0	70	0.05	0	0
255.6	71	-11.35	1.0	-11.35
272.9	69	5.95	-1.0	-5.95
평균	267.0	70.0		합계 -35.40
표준편차	8.2192	.8944		

공분산과 상관계수



▶ 표본 공분산

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{-35.40}{6-1} = \boxed{-7.08}$$

▶ 표본 상관계수

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-7.08}{(8.2192)(.8944)} = \boxed{-0.9631}$$