

2장, Part A

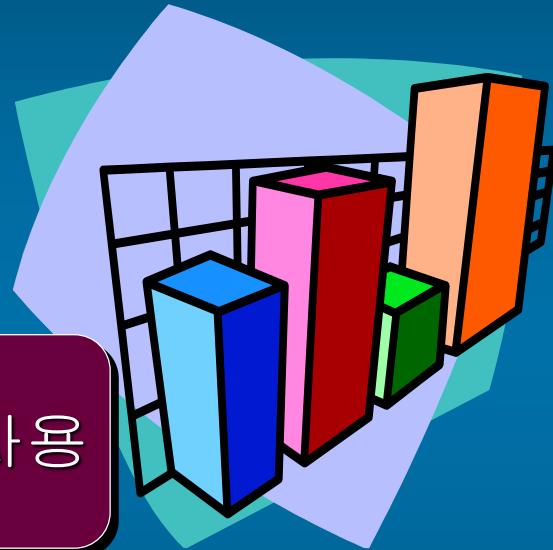
기술 통계:

표와 그래프로 나타내기

- 범주형 자료의 요약
- 양적 자료의 요약

범주형 자료는 라벨이나 이름을 사용

양적 자료는 수치 값을 사용



범주형(질적) 자료의 요약

- 도수분포
- 상대 도수분포
- 백분율 도수분포
- 막대 그래프
- 파이차트

도수분포(frequency distribution)

- ▶ 도수분포는 겹치지 않게 나눈 각 계급별 항목의 도수(개수)를 표로 요약해서 보여주는 것을 말한다
- ▶ 원 자료에서는 빨리 알아보기 힘든 (자료에 대한) 통찰력을 제공하는 것이 목적이다.

도수분포 예 : Marada Inn

Marada 여관에 투숙한 손님들은 숙박시설에 대하여 평가해줄 것을 요구 받는데, 평가 등급은 *excellent, above average, average, below average, Poor*이다. 20명의 표본 손님들에게서 받은 평가 내용이 아래와 같이 나타나 있다:



| | | |
|---------------|---------------|---------------|
| Below Average | Average | Above Average |
| Above Average | Above Average | Above Average |
| Above Average | Below Average | Below Average |
| Average | Poor | Poor |
| Above Average | Excellent | Above Average |
| Average | Above Average | Average |
| Above Average | Average | Average |

도수분포



| 등급 | 도수 |
|---------------|----------|
| Poor | 2 |
| Below Average | 3 |
| Average | 5 |
| Above Average | 9 |
| Excellent | <u>1</u> |
| 계 | 20 |

상대 도수분포(relative frequency distribution)

- ▶ 한 계급의 상대 도수는 그 계급에 속한 자료항목의 총수에 대한 분수표시나 비율이다.
- ▶ 상대 도수분포는 각 계급에 대한 상대도수를 보여주는 자료의 요약표이다.

백분율 도수분포(percent frequency distribution)

- ▶ 백분율도수는 상대도수에 100을 곱하면 된다.
- ▶ 백분율도수분포는 각 계급에 대한 백분율도수를 보여주는 자료의 요약표이다.

상대 도수와 백분율 도수 분포



| 등급 | 상대 도수 | 백분율도수 |
|---------------|------------|----------|
| Poor | .10 | 10 |
| Below Average | .15 | 15 |
| Average | .25 | 25 |
| Above Average | .45 | 45 |
| Excellent | <u>.05</u> | <u>5</u> |
| 계 | 1.00 | 100 |

$.10(100) = 10$

$1/20 = .05$

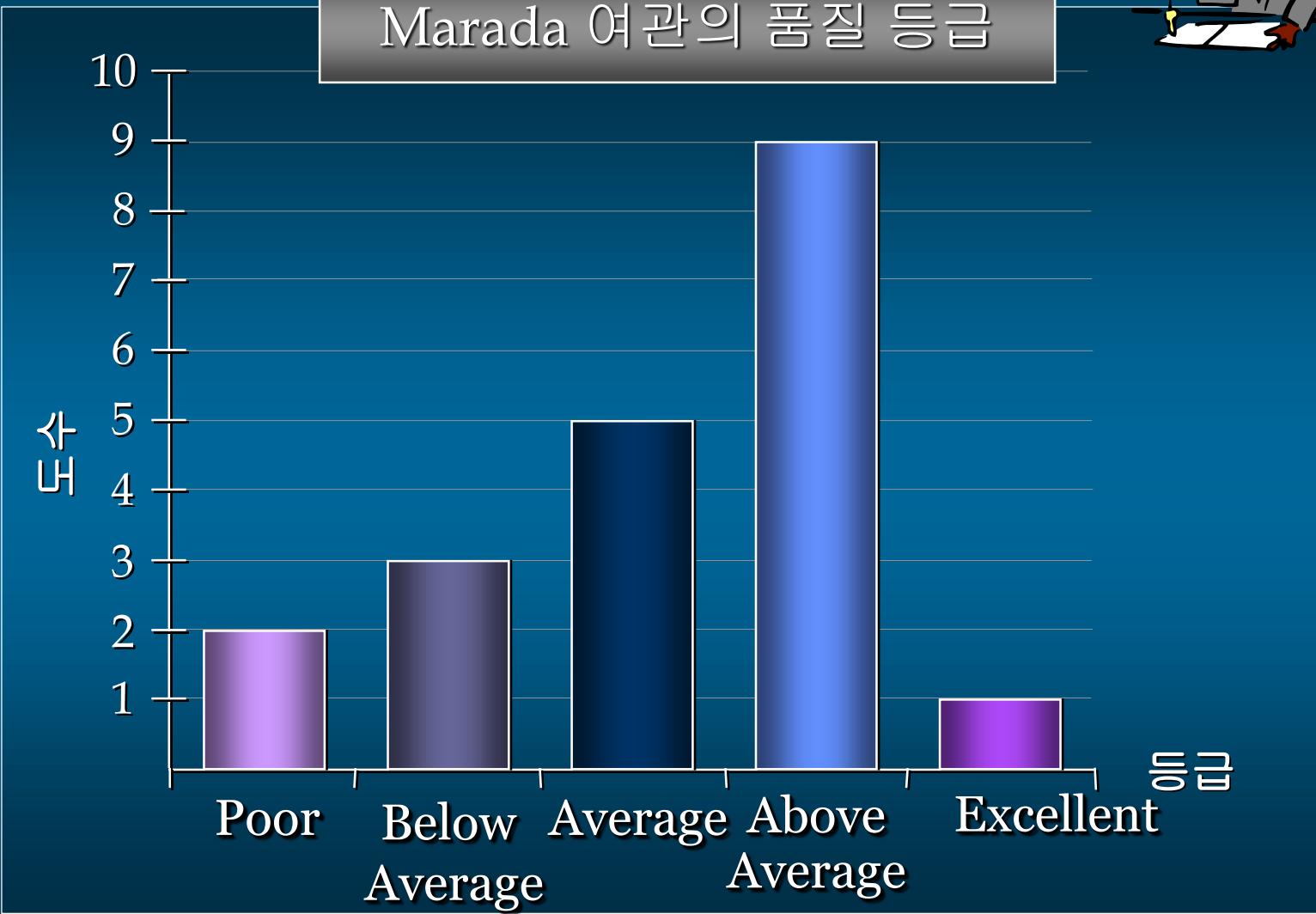
막대 그래프(bar graph)

- ▶ 막대 그래프는 질적자료를 묘사하는 그래프 도구이다.
- ▶ 그래프의 한 축(통상적으로 가로축)에 계급의 이름을 표기한다.
- ▶ 도수분포, 상대도수분포, 백분율도수분포의 크기(scale)는 그래프의 다른 한 축(보통 세로축)에 쓴다.
- ▶ 각 계급 이름 위에 고정 너비의 막대를 그리고, 도수에 따라 적절하게 막대의 길이를 늘려준다.
- ▶ 각 계급이 분리되어 있다는 것을 강조하기 위해서 막대는 서로 분리되어 있어야 한다

막대 그래프



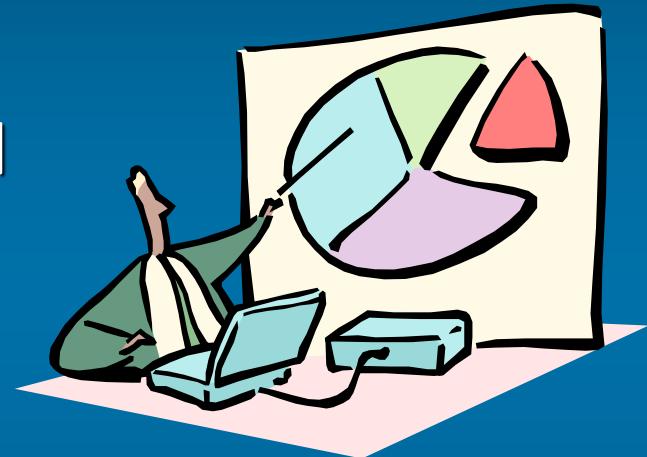
Marada 여관의 품질 등급



파이차트(pie chart)

▶ 파이차트는 질적 자료에 대한 상대 도수분포를 나타내기 위해 일반적으로 사용되는 그래프 도구이다.

▶ 먼저 원을 그린다. 그 원을 각 계급의 상대도수에 대응하는 면적 또는 부분으로 나눈다.

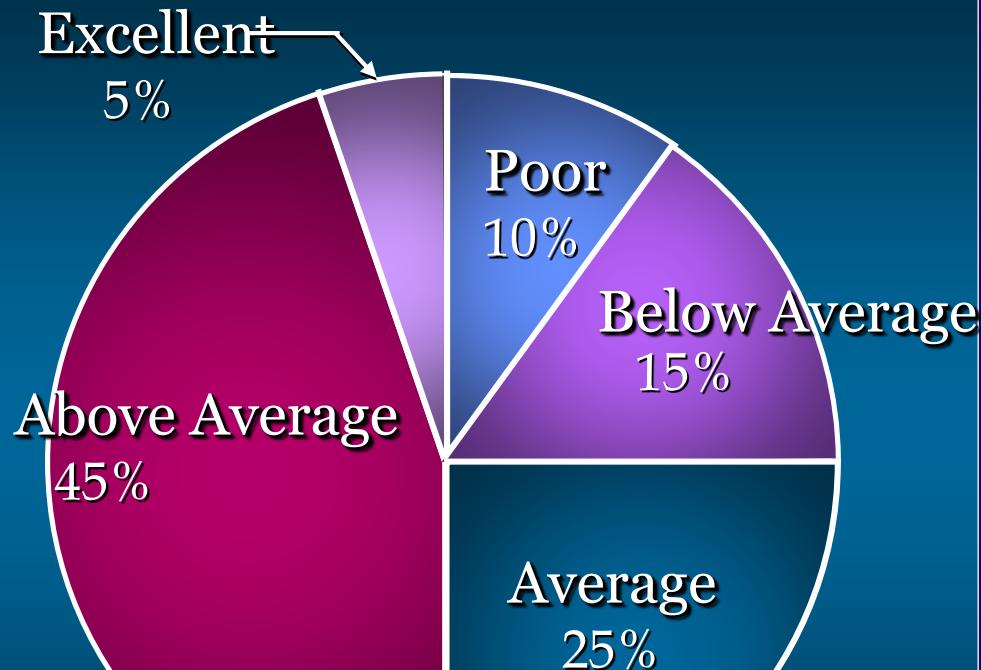


▶ 원은 360도 이므로, .25의 상대 도수를 가진 계급은 원의 $.25(360) = 90$ 도를 차지할 것이다.

파이차트



Marada 여관의 품질 등급



예 : Marada 여관



■ 파이차트로부터 얻은 통찰

- ▶ 조사된 고객의 반수가 Marada 여관의 품질 등급을 Above Average 또는 Excellent로 평가하였다(파이의 왼쪽 부분을 보면). 매니저가 기뻐할 것 같다.
- ▶ Excellent 등급을 준 고객 1명에 대하여 Poor 등급을 준 고객은 2명이다(파이의 위쪽 부분을 보면). 매니저 심기를 불편할 것이다.

양적 자료의 요약

- 도수분포
- 상대 도수분포와 백분율 도수분포
- 점그림
- 히스토 그램
- 누적 분포
- 누적 분포 그래프 (ogive)

Hudson 자동차 수리점

▶ Hudson 자동차 수리점의 관리자는 자신들의 수리점에서 행해지는 엔진조정에 사용되는 부품 비용에 대하여 더 자세한 조사를 하려 한다.



그래서 엔진조정을 한 50건의 고객송장을 조사하였다. 부품비용(달러 단위로 반올림 함)이 다음 슬라이더에 나타나 있다.

Hudson 자동차 수리점



- 50건의 엔진조정에 대한 부품비용 표본

| | | | | | | | | | |
|-----|----|----|-----|----|-----|----|----|----|-----|
| 91 | 78 | 93 | 57 | 75 | 52 | 99 | 80 | 97 | 62 |
| 71 | 69 | 72 | 89 | 66 | 75 | 79 | 75 | 72 | 76 |
| 104 | 74 | 62 | 68 | 97 | 105 | 77 | 65 | 80 | 109 |
| 85 | 97 | 88 | 68 | 83 | 68 | 71 | 69 | 67 | 74 |
| 62 | 82 | 98 | 101 | 79 | 105 | 79 | 69 | 62 | 73 |

도수분포

■ 계급수(number of classes)를 정하기 위한 지침

- ▶ 5개에서 20개 사이로 설정한다.
- ▶ 많은 요소(element)를 가진 자료집합은 보통 많은 계급수를 필요로 한다.
- ▶ 적은 수의 자료는 적은 계급수를 필요로 한다.

도수분포

■ 계급 크기(width of classes)를 정하기 위한 지침

- ▶ 계급의 크기는 동일하게 설정한다.
- ▶ 적정 계급 크기 =

$$\frac{\text{가장큰 자료값} - \text{가장작은 자료값}}{\text{계급의 수}}$$

도수분포



Hudson 자동차 수리점에서 6개의 계급을 정한다면:

▶ 적정 계급 크기 = $(109 - 52)/6 = 9.5 \cong 10$

| <u>부품비용 (\$)</u> | <u>도수</u> |
|------------------|-----------|
| 50-59 | 2 |
| 60-69 | 13 |
| 70-79 | 16 |
| 80-89 | 7 |
| 90-99 | 7 |
| 100-109 | <u>5</u> |
| 계 | 50 |

상대도수분포와 백분율도수분포



| 부품비용(\$) | 상대도수 | 백분율도수 |
|----------|------------|-----------|
| — | | |
| 50-59 | .04 | 4 |
| 60-69 | .26 | 26 |
| 70-79 | .32 | 32 |
| 80-89 | .14 | 14 |
| 90-99 | .14 | 14 |
| 100-109 | <u>.10</u> | <u>10</u> |
| 계 | 1.00 | 100 |

상대도수분포와 백분율도수분포



■ 백분율도수분포로부터 얻은 내용

- ▶ 부품비용의 4%만이 \$50-59 계급에 있다.
- ▶ 부품비용의 30%가 \$70 아래에 있다.
- ▶ 부품비용의 가장 많은 부분(32% 또는 거의 1/3)을 차지하는 계급은 \$70-79 계급이다.
- ▶ 부품비용의 10%는 \$100 이상에 있다.

점 그림(dot plot)

- 자료의 가장 단순한 도표형식 요약 중 하나가 점(dot) 그림이다.
- 수평축은 자료값의 범위를 나타낸다.
- 각각의 자료값은 수평축 위에 점으로 표현된다.

점 그림



엔진조정 부품 비용



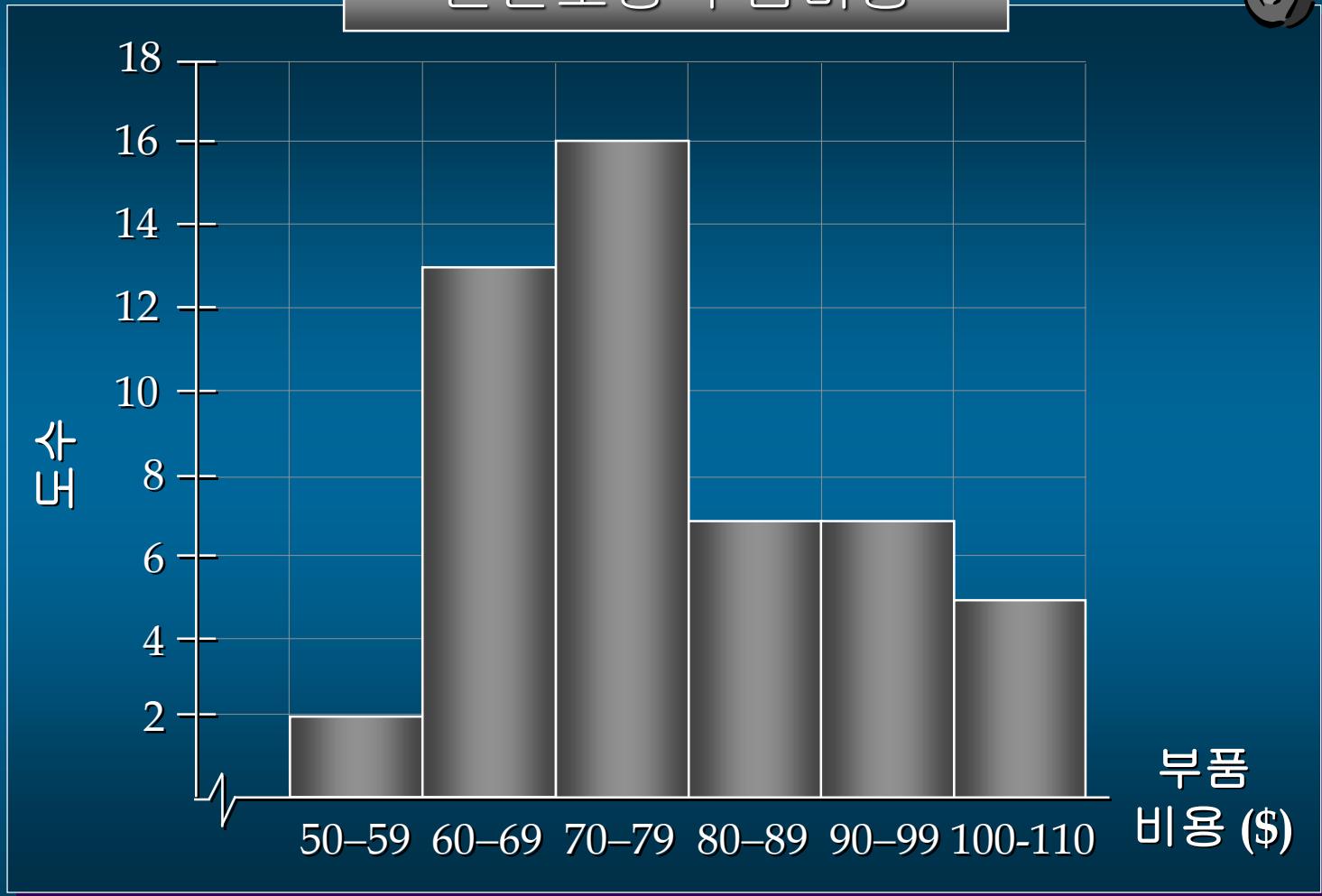
히스토그램(histogram)

- ▶ 양적 자료의 또 다른 그래프형식의 표현방식이 히스토그램이다.
- ▶ 관심의 대상이 되는 변수가 수평축 위에 놓인다.
- ▶ 각 계급구간 위에 사각형을 그리는데, 그 사각형의 높이는 각 계급구간에 해당하는 도수, 상대도수 또는 백분율상대도수를 기반으로 한다.
- ▶ 막대그래프와는 달리 히스토그램은 인접한 계급의 사각형끼리 구분이 없다.

히스토그램



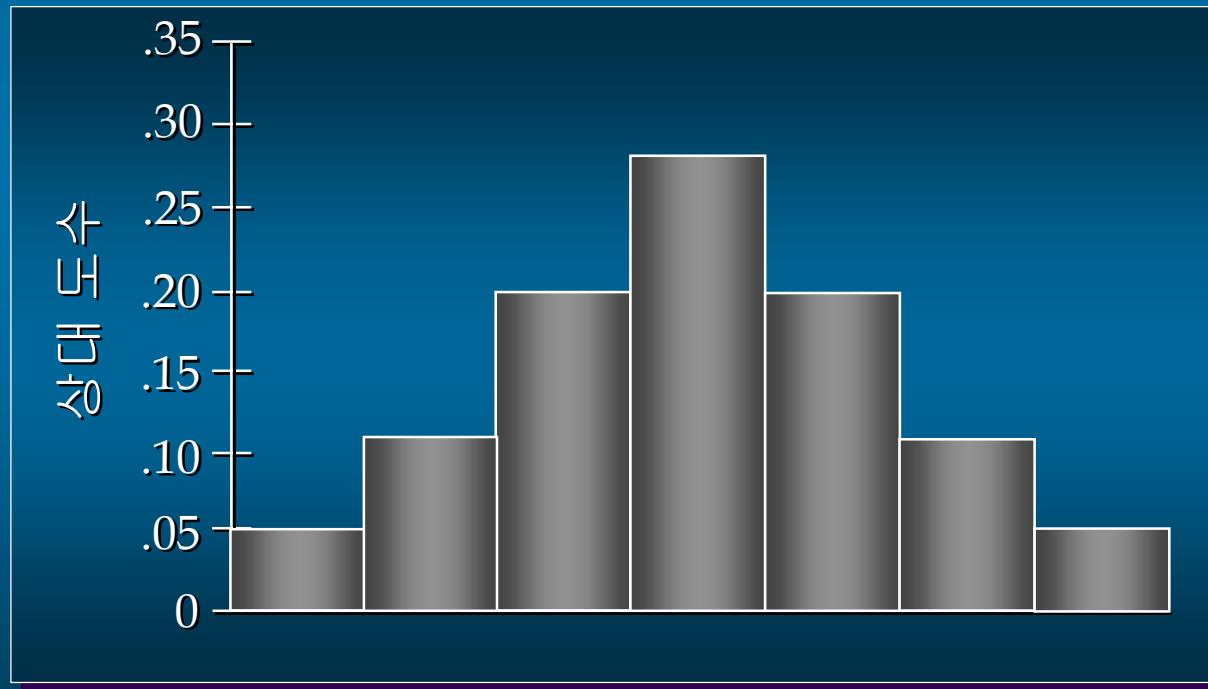
엔진조정 부품비용



히스토그램

■ 대칭 히스토그램

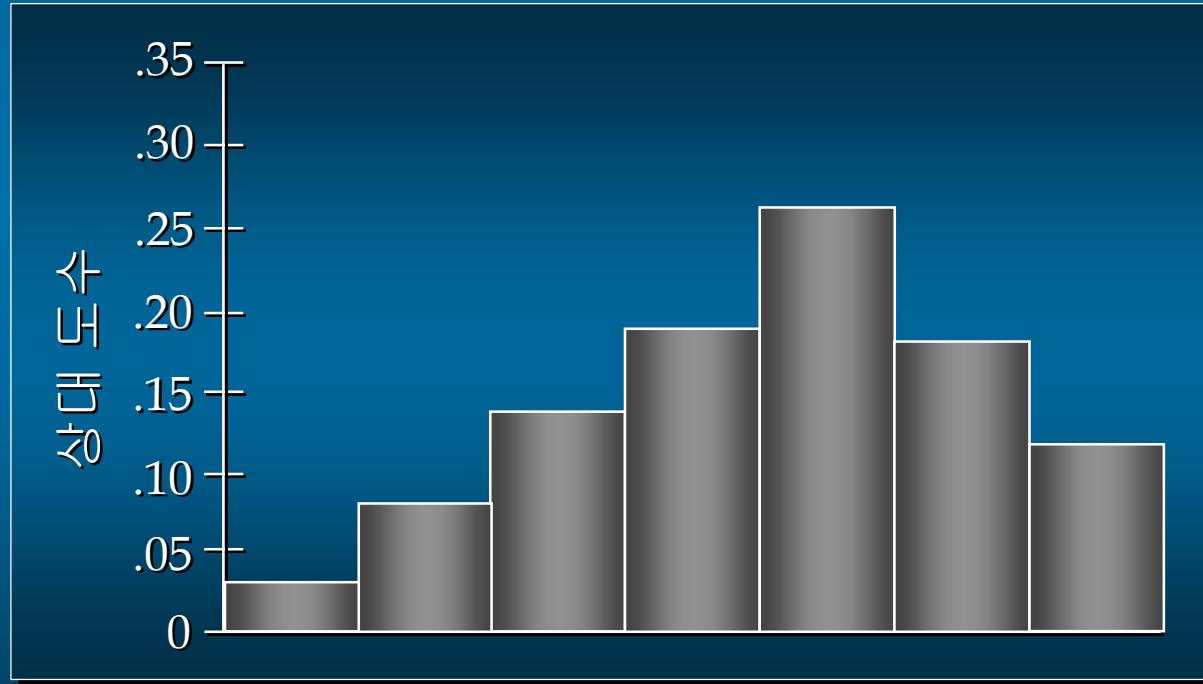
- 왼쪽꼬리의 모양이 거울에 오른쪽 꼬리가 비춰진 모양으로 나타난다.
- 예 : 사람들의 키와 몸무게



히스토그램

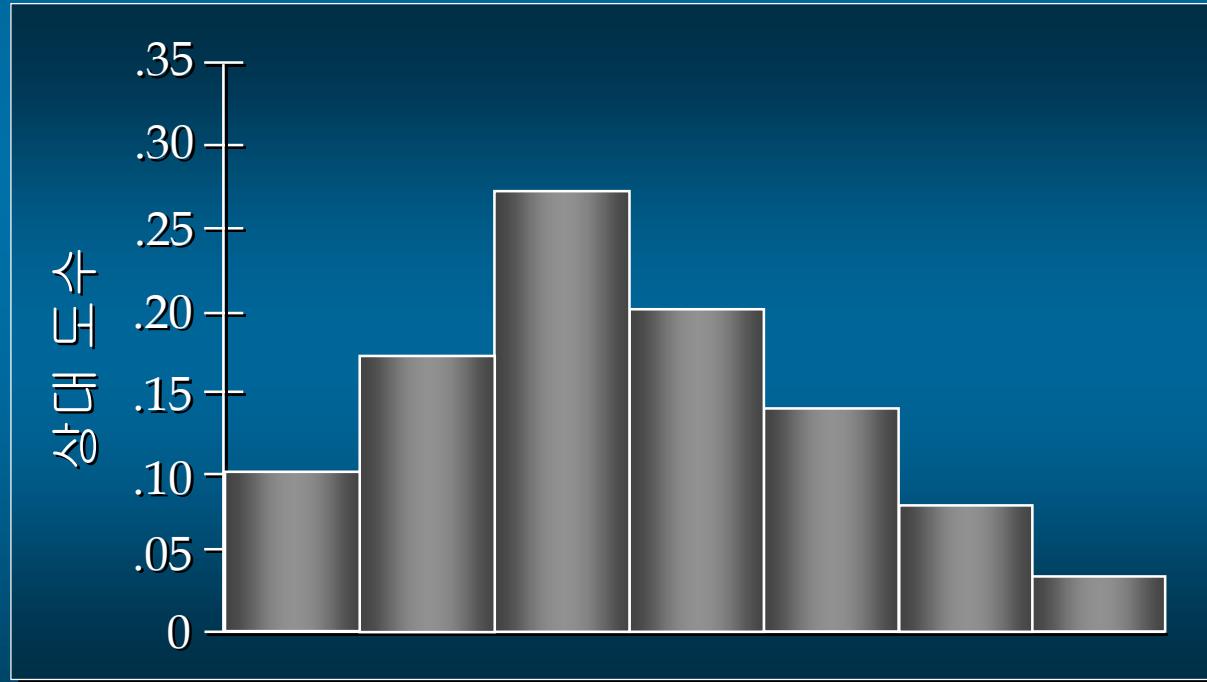
■ 적당히 왼쪽으로 경사진 히스토그램

- 왼쪽 꼬리가 더 길다
- 예 : 시험성적



히스토그램

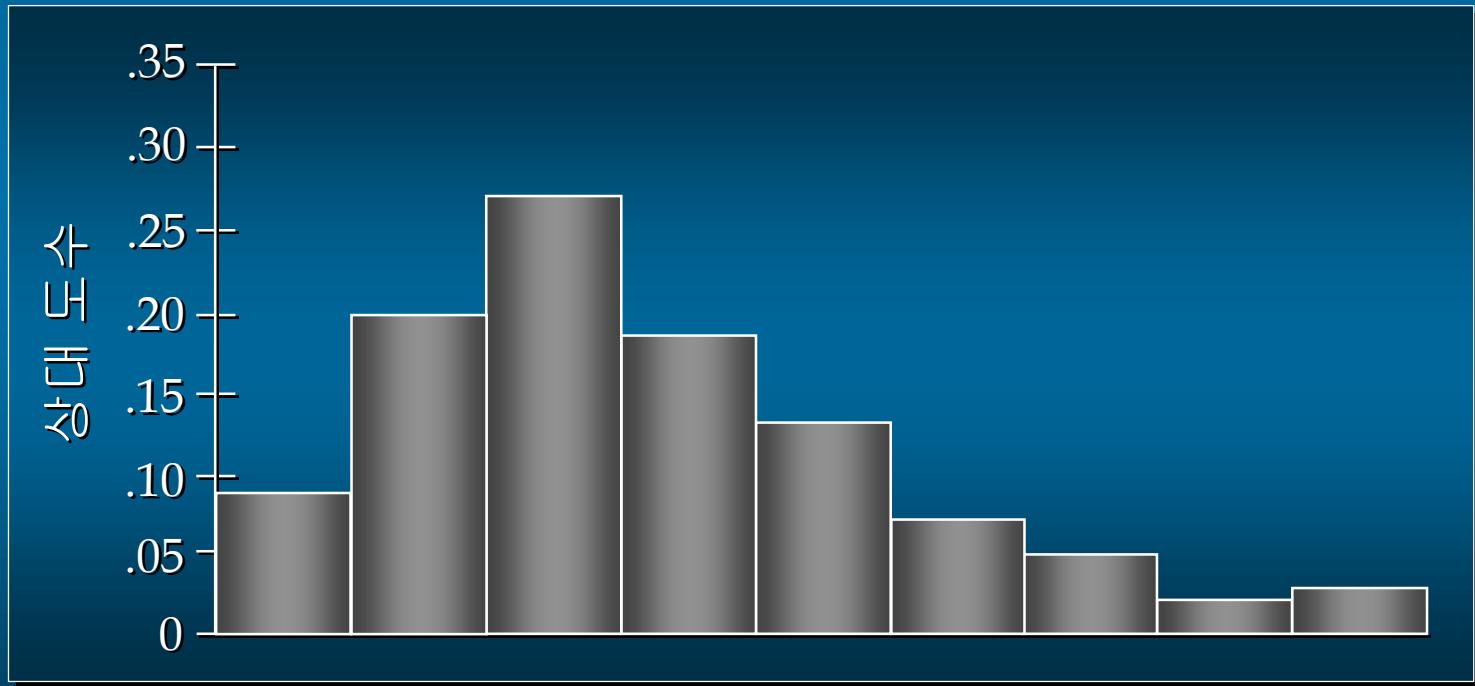
- 적당히 오른쪽으로 경사진 히스토그램
 - 오른쪽 꼬리가 더 길다.
 - 예 : 주택 가격



히스토그램

■ 오른쪽으로 심하게 경사진 히스토그램

- 매우 긴 오른쪽 꼬리
- 예 : 중역들의 급여



누적 분포(cumulative distribution)

- ▶ 누적 도수분포 – 각 계급의 상한값과 같거나 그 보다 작은 값을 가지는 항목의 수를 나타낸다.
- ▶ 누적 상대도수분포 – 각 계급의 상한값과 같거나 그 보다 작은 값을 가지는 항목의 비율을 보여준다.
- ▶ 누적 백분율도수분포 – 각 계급의 상한값과 같거나 그 보다 작은 값을 가지는 항목의 백분율을 보여준다.

누적 분포



■ Hudson 자동차 수리점

| 비용 (\$) | 누적 도수 | 누적 상대 도수 | 누적 백분율 도수 |
|------------|----------|----------------|-----------------|
| ≤ 59 | 2 | .04 | 4 |
| ≤ 69 | 15 | .30 | 30 |
| ≤ 79 | 31 | $2 + 13$ | $15/50$ |
| ≤ 89 | 38 | .62 | 62 |
| ≤ 99 | 45 | .76 | 76 |
| ≤ 109 | 50 | .90 | 90 |
| | | 1.00 | 100 |

누적분포그래프(ogive)

- ▶ 오자이브(ogive)는 누적분포의 그래프이다.
- ▶ 자료의 값은 수평축 위에 나타낸다.
- ▶ 수직축 위에 나타내는 것은:
 - 누적 도수 또는
 - 누적 상대 도수 또는
 - 누적 백분율 도수
- ▶ 각 계급의 도수는 점을 찍어 나타낸다.
- ▶ 찍힌 점들은 직선으로 연결된다.

누적분포그래프(Ogive)



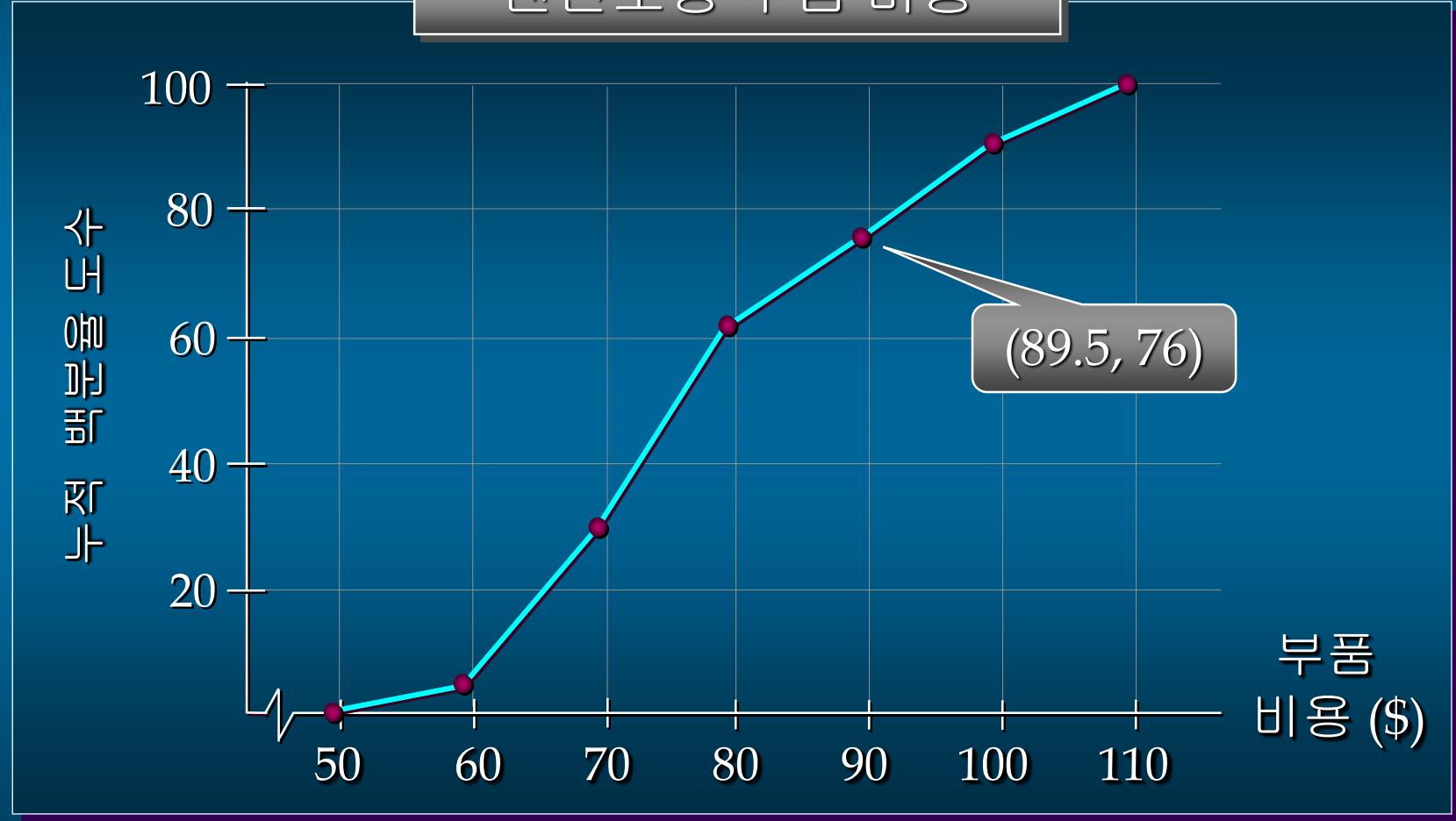
■ Hudson 자동차 수리점

- ▶ 부품비용 자료의 계급한계가 50-59, 60-69, 등등
이기 때문에, 59에서 60, 69에서 70, 등의 공백
구간이 나타난다.
- ▶ 이러한 공백은 각 계급한계의 중간값을 취함으로써
제거할 수 있다.
- ▶ 그래서, 59.5가 50-59 계급에 사용되고, 69.5가 60-
69계급에 사용된다.

누적백분율 도수의 오자이브



엔진조정 부품 비용

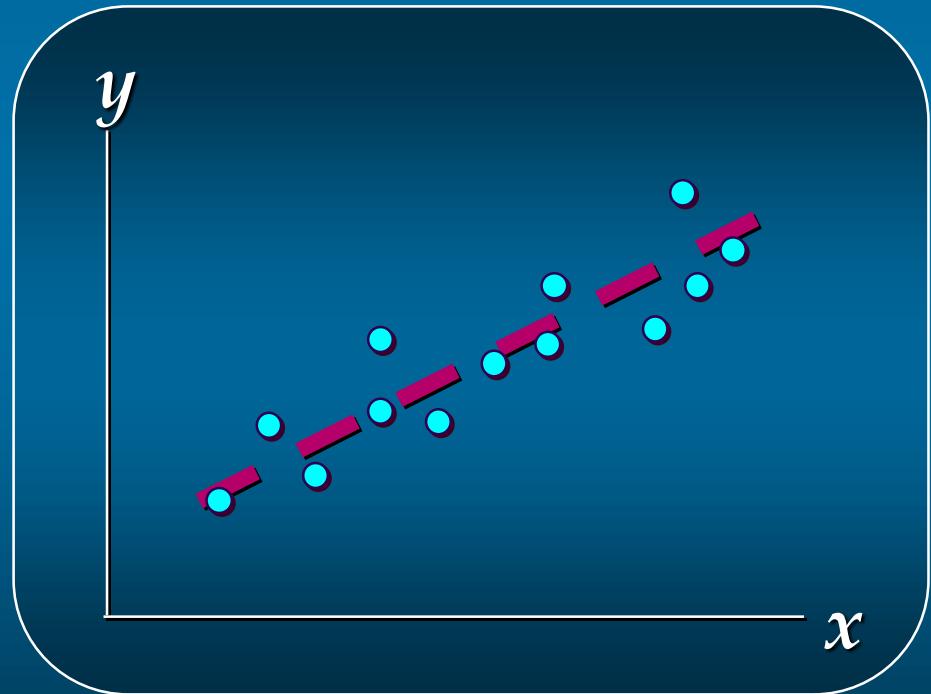


2장, Part B

기술 통계:

표와 그래프로 나타내기

- 탐색적 자료 분석
- 교차제표와 산점도



탐색적 자료 분석(exploratory data analysis)

- ▶ 탐색적 자료 분석의 기술은 간단한 산수와 쉽게 그릴 수 있는 그림으로 이루어져 있으며, 자료를 빠르게 요약하는데 사용된다.
- ▶ 그러한 하나의 기술은 줄기-잎 그림이다.

줄기-잎 그림(stem-and-leaf display)

- ▶ 줄기-잎 그림은 자료의 순위와 분포모양을 모두 보여 준다.
- ▶ 이런 측면에서 히스토그램과 비슷하지만, 실제 자료값을 보여주는 장점이 있다.
- ▶ 자료 항목의 앞 자리 숫자가 수직선의 왼쪽에 배열된다.
- ▶ 수직선의 오른편에는 각 자료의 마지막 자리 숫자를 순서대로 적는다.
- ▶ 그림의 각 행의 왼쪽에 있는 수를 줄기라 한다.
- ▶ 줄기에 있는 각 자리수는 잎이다.

예 : Hudson 자동차 수리점

▶ Hudson 자동차의 관리자는 자신들의 수리점에서 행해진 엔진 조정에 사용되는 부품비용에 대하여 좀더 자세한 조사를 하려 한다.

그래서 엔진조정을 한 50건의 고객 송장을 조사하였다.
부품비용(달러 단위로 반올림 함)이 다음 슬라이더에 나타나 있다.



예 : Hudson 자동차 수리점



- 50건의 엔진조정에 대한 부품비용 표본

| | | | | | | | | | |
|-----|----|----|-----|----|-----|----|----|----|-----|
| 91 | 78 | 93 | 57 | 75 | 52 | 99 | 80 | 97 | 62 |
| 71 | 69 | 72 | 89 | 66 | 75 | 79 | 75 | 72 | 76 |
| 104 | 74 | 62 | 68 | 97 | 105 | 77 | 65 | 80 | 109 |
| 85 | 97 | 88 | 68 | 83 | 68 | 71 | 69 | 67 | 74 |
| 62 | 82 | 98 | 101 | 79 | 105 | 79 | 69 | 62 | 73 |

줄기- 잎 그림



| | |
|----|---------------------------------|
| 5 | 2 7 |
| 6 | 2 2 2 2 5 6 7 8 8 8 9 9 9 |
| 7 | 1 1 2 2 3 4 4 5 5 5 6 7 8 9 9 9 |
| 8 | 0 0 2 3 5 8 9 |
| 9 | 1 3 7 7 7 8 9 |
| 10 | 1 4 5 5 9 |

줄기

잎

자세한 줄기-앞 그림

- ▶ 만약 줄기-앞 그림이 자료를 너무 간략하게 나타냈다고 생각한다면, 각 앞 자리수마다 두 개 이상의 줄기를 사용해서 그림을 더 펼칠 수 있다.
- ▶ 같은 줄기가 두 번 표시될 때는, 첫 번째 줄기는 0에서 4의 앞을 표시하고, 두 번째 줄기는 5에서 9의 앞을 표시한다.

자세한 줄기- 일 그림



| | |
|----|-------------------|
| 5 | 2 |
| 5 | 7 |
| 6 | 2 2 2 2 |
| 6 | 5 6 7 8 8 8 9 9 9 |
| 7 | 1 1 2 2 3 4 4 |
| 7 | 5 5 5 6 7 8 9 9 9 |
| 8 | 0 0 2 3 |
| 8 | 5 8 9 |
| 9 | 1 3 |
| 9 | 7 7 7 8 9 |
| 10 | 1 4 |
| 10 | 5 5 9 |

줄기-잎 그림

- 잎 단위
- ▶ 각 잎을 구성하는 데 한 자리 자릿수가 쓰였다.
- ▶ 앞의 예에서는 잎 단위는 10이다.
- ▶ 잎 단위는 100, 10, 1, 0.1 등일 수 있다.
- ▶ 줄기-잎 그림에서 잎의 단위가 나타나 있지 않다면, 잎 단위는 10이라고 가정한다.

예 : 잎 단위 = 0.1

다음과 같은 자료를 가지고 있다면:

8.6 11.7 9.4 9.1 10.2 11.0 8.8

이 자료의 줄기-잎 그림은:

잎 단위 = 0.1

| | | |
|----|---|---|
| 8 | 6 | 8 |
| 9 | 1 | 4 |
| 10 | 2 | |
| 11 | 0 | 7 |

예 : 일 단위 = 10

다음과 같은 자료를 가지고 있다면:

1806 1717 1974 1791 1682 1910 1838

이 자료의 줄기-일 그림은:

일 단위 = 10

| | |
|----|-----|
| 16 | 8 |
| 17 | 1 9 |
| 18 | 0 3 |
| 19 | 1 7 |

1682에는 82이나,
80으로 내림 되어
8로 표시 되었다.

교차제표와 산점도 (crosstabulations and scatter diagrams)

- ▶ 지금까지 우리는 한 번에 하나의 변수에 대해 자료를 요약하기 위한 표와 도표 형식의 방법들에 초점을 두었다.
- ▶ 종종 관리자나 의사결정권자는 두 변수들 사이의 관계를 이해하는데 도움이 되는 표 형식과 도표 형식의 방법들을 필요로 한다.
- ▶ 교차제표와 산점도가 바로 동시에 두 변수(이상)에 관한 자료를 요약 할 수 있는 방법들이다.

교차제표

- ▶ 교차제표는 두 변수에 대한 자료를 표 형식으로 요약한 것이다.
- ▶ 교체제표는 다음과 같은 상황에서 사용된다
 - 한 변수는 질적 변수이고, 다른 변수는 양적 변수일 때
 - 두 변수 모두 질적 변수 이거나,
 - 두 변수 모두 양적 변수 일 때
- ▶ 왼쪽과 상단의 이름표는 두 변수에 대한 등급을 정의한다.

교차제표



■ 예 : Finger Lakes Homes

지난 2년 동안 판매된 Finger Lakes 주택을 스타일과 가격으로 분류하여 아래에 표시 하였다.

| 가격대 | 주택 스타일 | | | | 계 |
|------------|--------|-----|-----|-------|-----|
| | 콜로니얼 | 통나무 | 스프릿 | A-프레임 | |
| ≤ \$99,000 | 18 | 6 | 19 | 12 | 55 |
| > \$99,000 | 12 | 14 | 16 | 3 | 45 |
| 계 | 30 | 20 | 35 | 15 | 100 |

교차제표



■ 앞의 교차제표에서 얻은 내용

- ▶ 가장 많은 수를 차지하는 부분은 (19) 스프릿 스타일에 \$99,000보다 작거나 같은 가격의 주택이다 .
- ▶ A-프레임 스타일과 동시에 \$99,000 보다 큰 가격에 속하는 주택은 3채이다.

교차제표



가격변수에 대한 도수 분포

| 가격대 | 주택 스타일 | | | | 계 |
|------------|--------|-----|-----|-------|-----|
| | 콜로니얼 | 통나무 | 스프릿 | A-프레임 | |
| ≤ \$99,000 | 18 | 6 | 19 | 12 | 55 |
| > \$99,000 | 12 | 14 | 16 | 3 | 45 |
| 계 | 30 | 20 | 35 | 15 | 100 |

주택 스타일에 대한 도수 분포

교차제표: 행 또는 열 백분율

- 교차제표에 기입되어 있는 값을 행 백분율 또는 열 백분율로 치환하면 두 변수 간의 관계를 더 잘 파악할 수 있다.

교차제표 : 행 백분율



| 가격대 | 주택 스타일 | | | | | 계 |
|------------|--------|-------|-------|-------|-----|---|
| | 콜로니얼 | 통나무 | 스프릿 | A-프레임 | | |
| ≤ \$99,000 | 32.73 | 10.91 | 34.55 | 21.82 | 100 | |
| > \$99,000 | 26.67 | 31.11 | 35.56 | 6.67 | 100 | |

주: 열의 총계는 반올림 때문에 100.01 이 된다.

$$(콜로니얼 > \$99K)/(모든 주택 >\$99K) \times 100 = (12/45) \times 100$$

교차제표: 열 백분율



| 가격대 | 주택 스타일 | | | |
|------------|--------|-------|-------|-------|
| | 콜로니얼 | 통나무 | 스프릿 | A-프레임 |
| ≤ \$99,000 | 60.00 | 30.00 | 54.29 | 80.00 |
| > \$99,000 | 40.00 | 70.00 | 45.71 | 20.00 |
| 계 | 100 | 100 | 100 | 100 |



$$(\text{콜로니얼} > \$99K) / (\text{모든 콜로니얼}) \times 100 = (12/30) \times 100$$

교차제표: 심슨의 역설(Simpson's paradox)

- ▶ 두 개 혹은 그 이상의 교차제표에 나온 자료를 합쳐 요약 교차제표를 만들기도 한다.
- ▶ 종합(aggregated) 교차제표에 있는 두 변수 사이의 관계에 대해는 주의해서 결론을 도출해야 한다.
- ▶ 심슨의 역설: 어떤 경우에는 각각의 자료를 따로 봤을 때와 종합교차제표를 보았을 때 정반대의 결론이 나오기도 한다.

종합교차제표



| 평결 | 판사 | | 계 |
|----|------------|------------|-----|
| | Luckett | Kendall | |
| 유지 | 129 (86%) | 110 (88%) | 239 |
| 번복 | 21 (14%) | 15(12%) | 36 |
| 계 | 150 (100%) | 125 (100%) | 275 |

종합교차제표



평결

Luckett 판사

민사법원

지방법원

계

유지
번복

29 (91%)

100 (85%)

129

3 (9%)

18(15%)

21

계

32 (100%)

118 (100%)

150

평결

Kendall 판사

민사법원

지방법원

계

유지
번복

90 (90%)

20 (80%)

110

10 (10%)

5 (20%)

15

계

100 (100%)

25 (100%)

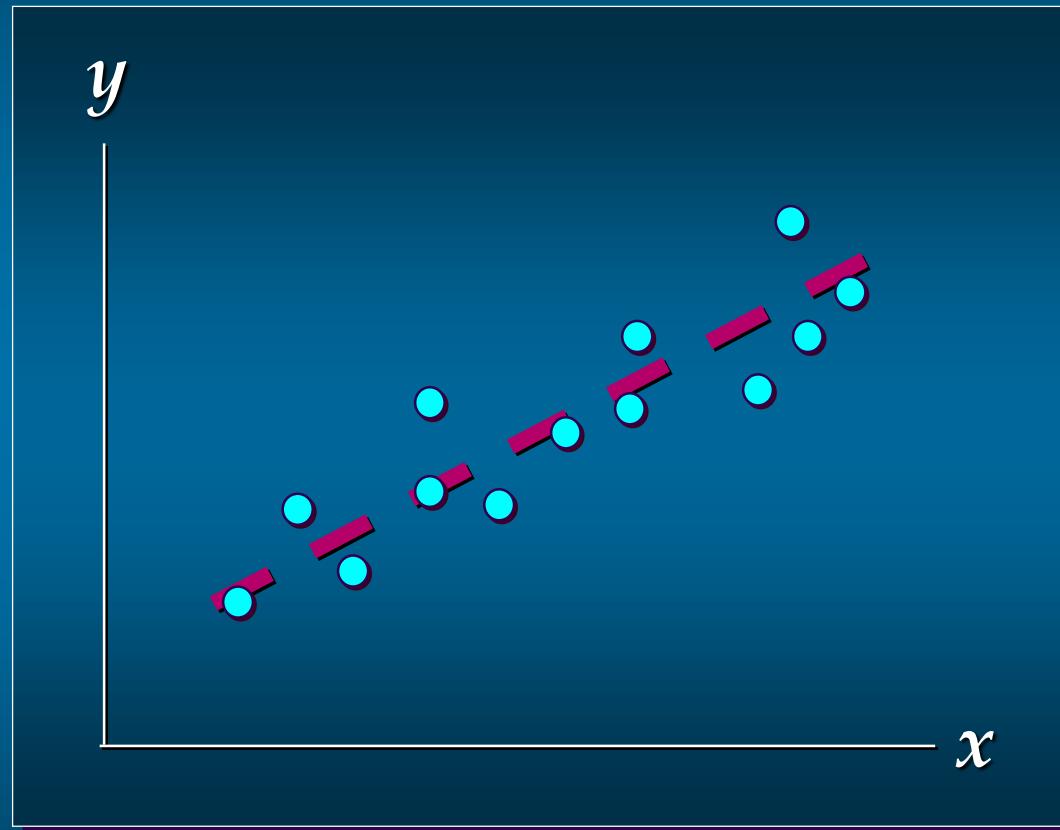
125

산점도와 추세선(scatter diagram and trendline)

- ▶ 산점도는 두 양적변수 사이의 관계를 그래프로 표현한 것이다.
- ▶ 한 변수는 가로축에 나타나고 다른 변수는 세로축에 나타난다.
- ▶ 표시된 점들의 일반적인 패턴은 변수들간의 전반적인 관계를 나타낸다.
- ▶ 추세선은 그 관계의 근사값을 보여주는 선이다.

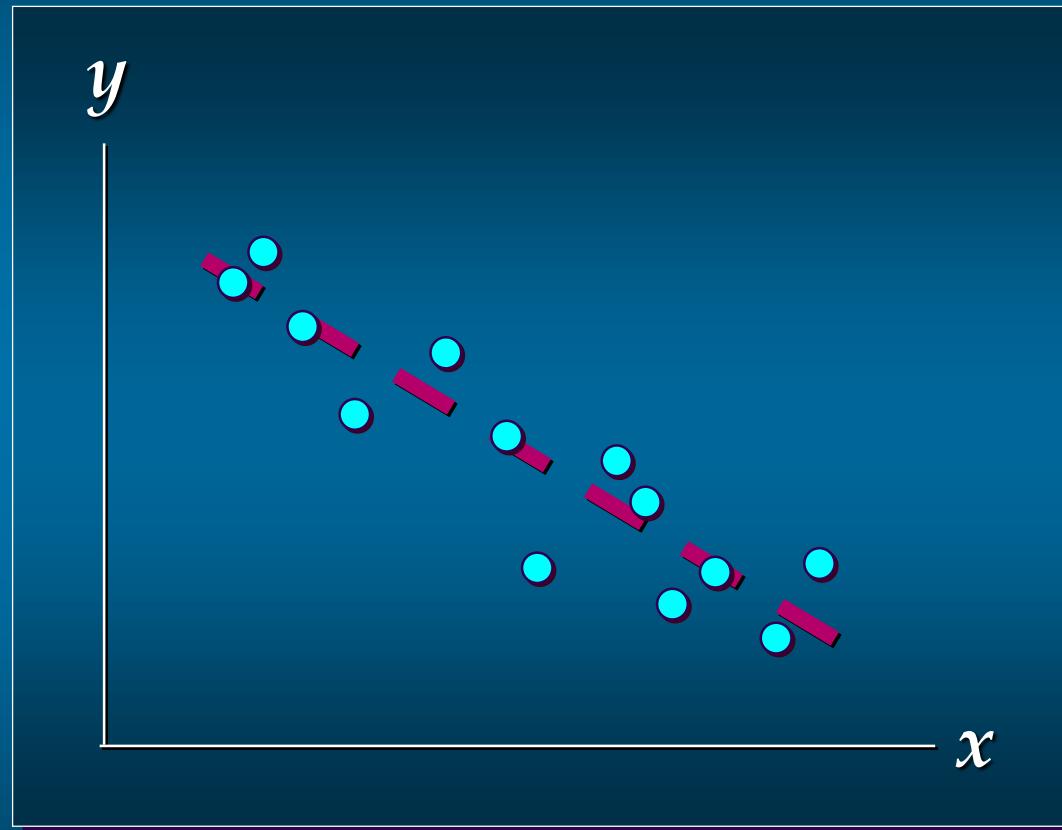
산점도

■ 정의 관계



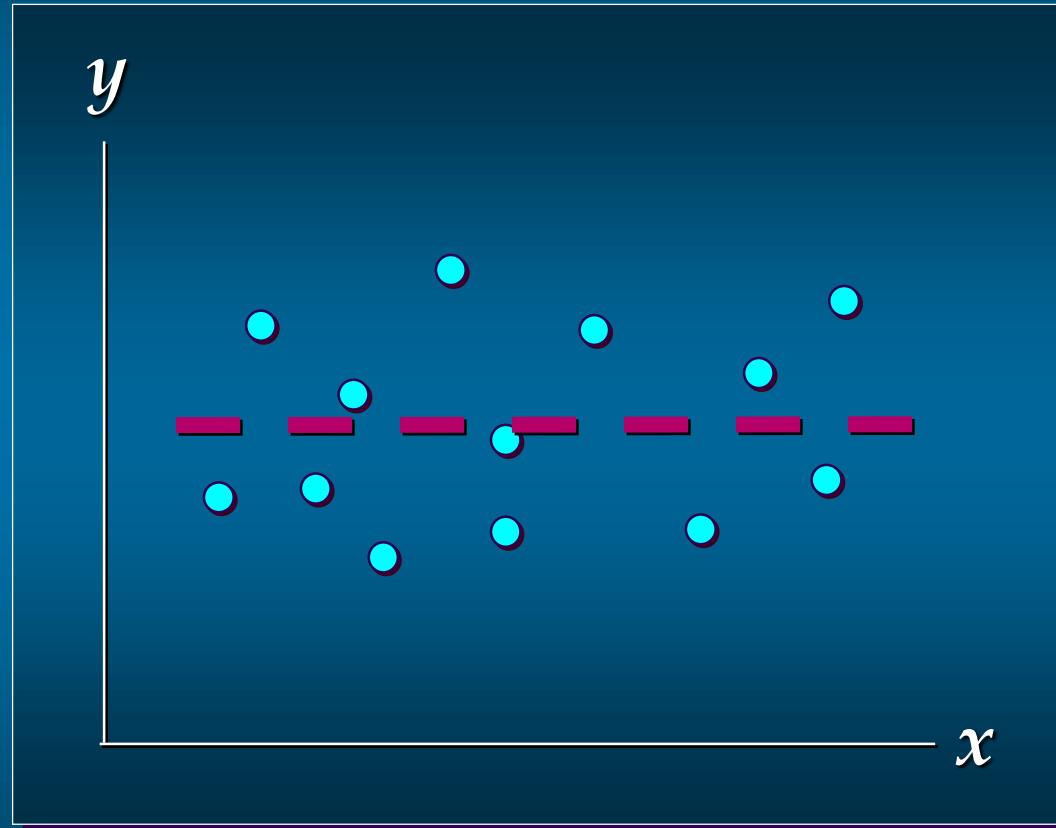
산점도

■ 역의 관계



산점도

■ 뚜렷한 관계가 없는 경우



예 : Panthers Football Team

■ 산점도

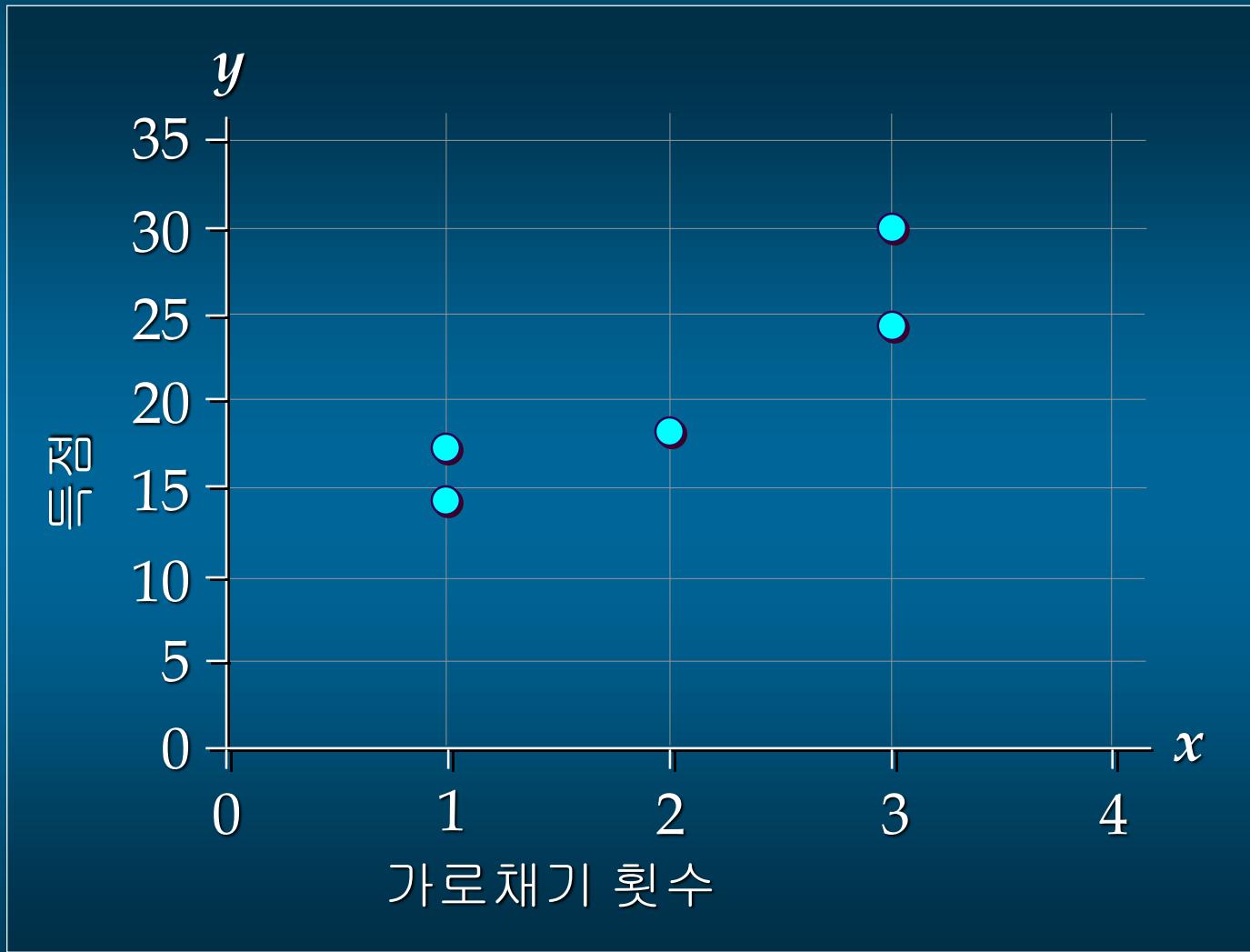
Panthers football team은 가로채기 횟수와 득점과의 관계에 대하여 조사하고 싶어 한다



$$x = \text{가로채기 횟수} \quad y = \text{득점}$$

| | |
|---|----|
| 1 | 14 |
| 3 | 24 |
| 2 | 18 |
| 1 | 17 |
| 3 | 30 |

산점도



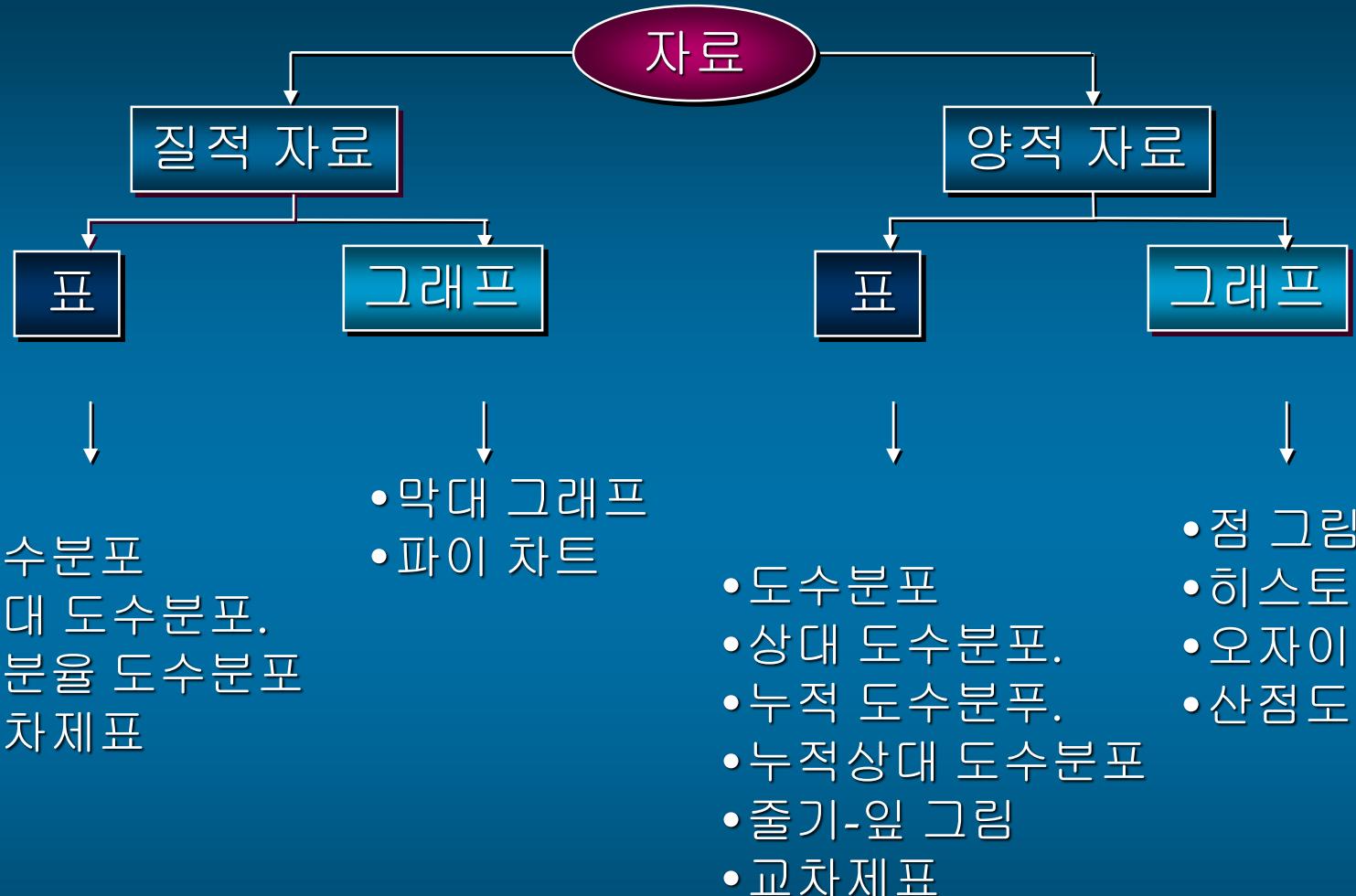
예 : Panthers Football Team



■ 앞의 산점도로부터 얻은 내용

- ▶ 앞의 산점도는 가로채기 횟수와 득점간에 정의관계를 보여 준다.
- ▶ 높은 득점은 높은 횟수의 가로채기와 관련되어 있다.
- ▶ 완벽한 정의 관계는 아니다. 모든 점이 일직선상에 나타나지 않는다.

표와 그래프의 절차



2장, Part B 끝

