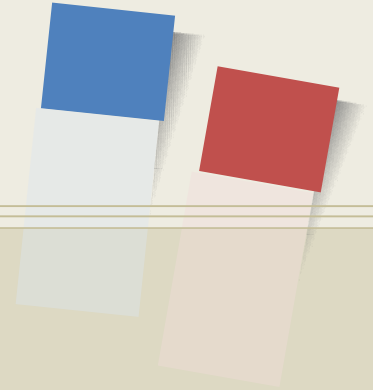


2020년 2학기

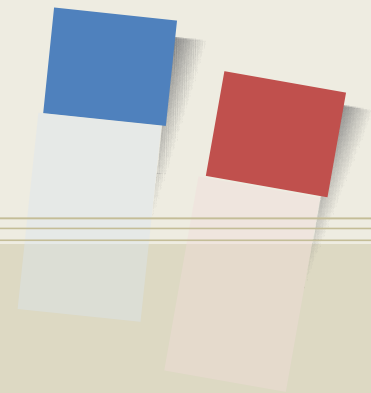


# 1주1강. 강의개요 소개, 자료의 표현1

확률 및 통계



## 강의개요 소개





## 담당교수: 이혜숙



- ❑ 연락처 : 010-7358-2165
- ❑ E-mail : [hyesook@knu.ac.kr](mailto:hyesook@knu.ac.kr)
- ❑ 면담시간 : 수업시간 전후(권장)  
사전연락 후 언제나

# 수강 참고사항

## □ 코로나 관련 공지사항(1주~2주)

1. 개강 후 2주간은 강의지원시스템을 통한 비대면 강의로 진행합니다.
2. 비대면 강의방식 :
  1. 동영상+수업자료+과제
  2. 동영상은 ppt파일에 음성지원을 기본으로 하며, 효율성이 좋은 것으로 전환이 될 수도 있습니다.
3. 코로나로 인해 강의와 시험의 일정과 유형이 바뀔 수 있습니다.

# 수강 참고사항

## □ 코로나 관련 공지사항(3주~)

1. 코로나가 안정이 된다면 대면으로 수업이 진행될 예정입니다.
2. 코로나 상황에 따라 비대면으로 전환될 수 있습니다.

# 수강 참고사항

## □ 일반 공지사항

1. 시험은 과제로 주어진 연습문제에서 낼 예정이므로 매주 수업후 반드시 풀어 볼 것.
2. 과제는 추후에 공지가 됩니다. 공지에 따라 제출하시기 바랍니다.

# 과제 및 평가방법

- ❑ 중간고사 : 40점 만점으로 실시
- ❑ 기말고사 : 40점 만점으로 실시
  - ❑ 형식 : 서술형 문제 등
- ❑ 과제물 및 출석(각 10점)
  - ❑ 출석 : 대면출석과 비대면 출석에 따라 10점 만점으로 평가함.  
(결석 2시간에 1점 감점을 기본으로 함)
  - ❑ 과제 : 과제물 제출에 따른 10점 만점으로 평가함.
- ❑ 과목 공동시험, 공동평가
  - ❑ 중간고사 : 8주차(예정)
  - ❑ 기말고사 : 15주차(예정)

# 강의교재 및 참고문헌

## 통계학의 이해(8판)

- 저자 : 이용구, 김삼용
- 출판사 : 율곡출판사
- 한글 번역판

## 참고 문헌

- B. Engelhardt, Introduction to probability and mathematical statistics, Duxbury Press
- Roy D. Yates and David J, Probability and Random Processes for Electrical Engineering, 2nd ed.

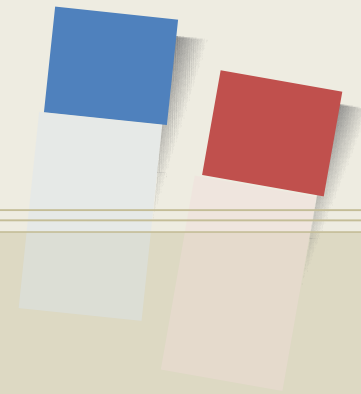




강의계획	강의주제	강의내용	과제	평가
1주차	자료의 표현	자료의 숫자요약	연습문제	
2주차	확률	베이지즈정리	연습문제	
3주차	확률과분포	확률변수 등	연습문제	
4주차	확률과분포	평균,분산 등	연습문제	
5주차	이산형 확률분포	베르누이분포	연습문제	
6주차	이산형 확률분포	다항분포	연습문제	
7주차	연속형 확률분포	균등분포	연습문제	
8주차	연속형 확률분포	지수분포	연습문제	중간 시험

9주차	확률표본과 추정	확률표본과 통계량, T-분포 등	연습문제	
10주차	가설검정	기초개념	연습문제	
11주차	가설검정	F-검정 등	연습문제	
12주차	가설검정	유의수준 등	연습문제	
13주차	분산분석	일원분류 분산분석	연습문제	
14주차	회귀분석과 상관 분석	단순형 회귀분석 등	연습문제	
15주차	빅데이터와 통계학	빅데이터와 통계학	연습문제	기말 시험

2.1절~2.3절

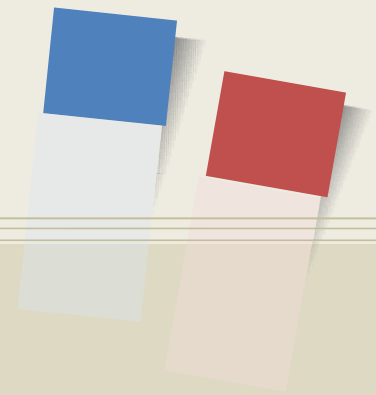


# 자료의 표현1





## 2.1절. 통계분석과 자료수집



## 통계분석과 통계적 추론

- **통계분석** : 특정 집단을 대상으로 자료를 수집해 대상집단에 대한 정보를 구하고, 적절한 통계분석 방법을 이용해 의사결정(통계적 추론)을 하는 과정.
- **통계적 추론** : 수집된 자료를 이용해 대상집단(모집단)에 대해 의사결정 하는 것.
  - ▣ **추정** : 대상집단의 특성값(모수)이 무엇일까?"를 추측
  - ▣ **가설검정** : 대상집단에 대해 특정한 가설을 설정한 후 그 가설의 채택여부를 결정

## 사례1--표본조사

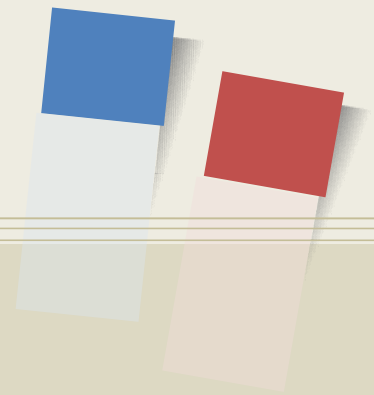
- 한 여론조사기관에서 1997년 12월에 국회에서 의결한 금융실명제 보완 입법에 대한 국민들의 지지율을 조사했다.
  - ▣ 구하고자 하는 정보 - 금융실명제 보완 입법에 대한 국민들의 지지율
  - ▣ 대상집단 - 우리 나라 유권자
  - ▣ 특성값(모수) - 지지율
  - ▣ 표본이 대상집단을 잘 대표할 수 있도록 추출되어야 함

## 사례2--실험

- ❑ 한 제약회사에서 새로 개발된 AIDS 치료제의 효과를 분석하는 실험을 실시.
  - ❑ 실험방법 1  
AIDS 에 감염된 환자 20 명을 임의로 선발하여 위의 치료제를 투약한 후,  
시간의 흐름에 따른 치료 효과를 측정
  - ❑ 실험방법 2  
AIDS 에 감염된 환자 20 명을 랜덤하게 10 명씩 두 집단으로 나눈 후,  
한 집단에는 새로 개발된 치료제를 투약하고,  
다른 집단에는 치료제를 투약하지 않은 채 시간의 흐름에 따른 두 집단의 반응을 비교



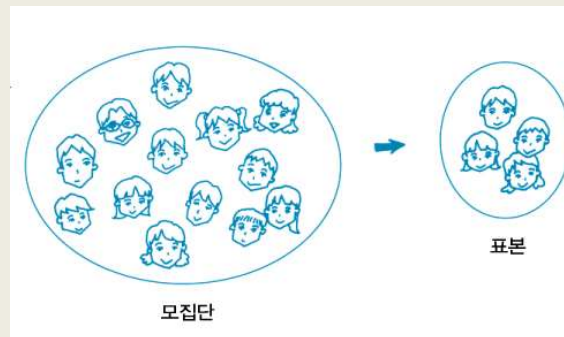
## 2.2절. 표본조사





# 총 조사, 표본조사

- ❑ **총조사** : 대상집단 모두를 조사하는 방법. '인구 및 주택 총조사'가 그 대표적인 예.
  - ▣ 많은 시간과 비용이 소요됨.
- ❑ **표본조사** : 대상집단의 일부를 관측해 그 대상집단 전체에 대한 정보를 구하는 과정.
  - ▣ 총조사에 비해 시간과 비용을 절약할 수 있음. (특히 파괴검사에 유용)



## 용어

- ❑ **모집단** : 조사하고자 하는 대상집단전체
- ❑ **원소** : 모집단을 구성하는 최소 단위
- ❑ **표본** : 조사하기 위하여 뽑힌 모집단의 일부
- ❑ **모수** : 표본관측에 의하여 구하고자 하는 모집단의 특성치

# 표본조사 시 유의사항

## 1. 표본이 합리적으로 추출되었는가?

- ▣ 이 표본이 모집단을 잘 대표할 수 있는가?
- ▣ 연령계층별, 출신지역별, 성별, 학력별, 직업별, 기타 사회계층별

## 2. 질문의 형식이 특정사항을 선호하도록 유도 금지.

- ▣ 낙태에 대한 찬성률을 조사하는데 질문을 "당신은 낙태를 합법화하는데 찬성하십니까?"로 묻는 경우와 "태아를 하나의 생명체로 존중해야 합니까?"로 묻는 경우

## 3. 조사방법에 따라서도 조사결과에 차이가 있을 수 있다.

- ▣ 면접조사, 우편조사, 전화조사(시간고려)

## 4. 표본조사 시점에 대한 고려를 해야 한다.

- ▣ 특정 사건 후 주의

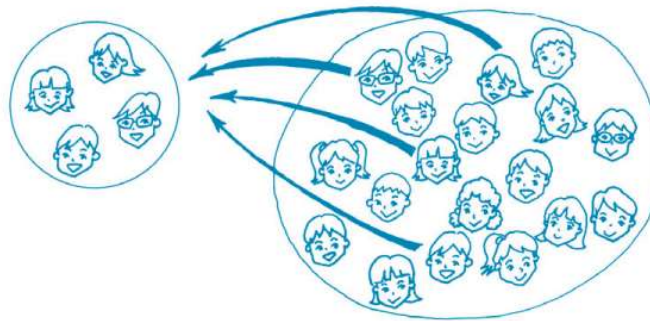
## 5. 표본조사의 결과에 대한 절대적인 신뢰를 해서는 안 된다.

# 표본추출방법

## (1) 단순랜덤추출법

- 모집단의 각 원소에  $1, 2, 3, \dots, N$ 까지의 번호를 부여하고, 그 중에서  $n$ 개의 번호를 임의로 선택하여 그 번호에 해당하는 원소를 표본으로 추출하는 방법.

그림 2-1 단순랜덤추출법

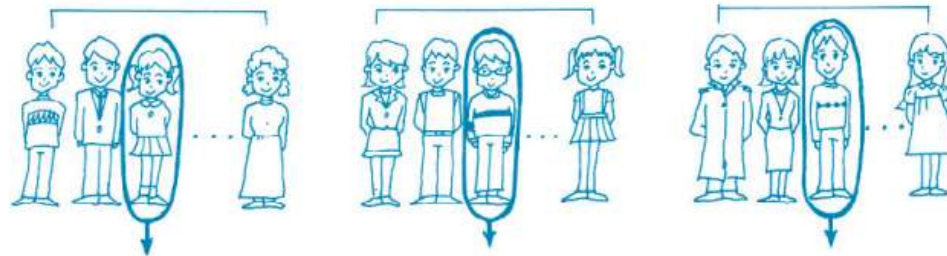


# 표본추출방법

## (2) 계통추출법

- 모집단의 모든 원소들에게  $1, 2, 3, \dots, N$ 의 일련번호를 부여하고 이를 순서대로 나열한 후에  $k$ 개 씩  $n$ 개의 구간으로 나누고 첫 구간( $1, 2, \dots, k$ )에서 하나를 임의로 선택한 후에  $k$ 개씩 띄어서 표본을 추출하는 방법.

그림 2-2 계통추출법



# 표본추출방법

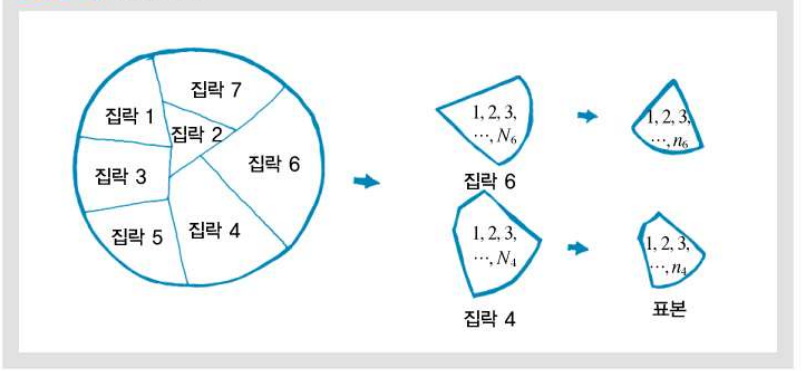
## (3) 집락추출법

- ❑ 모집단이 몇 개의 집단이 결합된 형태로 구성되어 있고, 각 집단 내부에서는 원소들에게 일련번호를 부여할 수 있는 경우에 이용되는 표본추출방법.
- ❑ 각 집단을 집락(cluster)이라고 하는데 표본추출과정은 일부 집락을 랜덤으로 선택하고 선택된 각 집락 내에서 표본을 임의로 선택하는 방법

예) 서울시내의 가구를 조사대상으로 조사하는 경우

- ▷ 서울시 25개 구 중 5개 구 추출
- ▷ 추출된 5개 구에서 임의로 4개 동 추출
- ▷ 추출된 동에서 50개 가구 추출

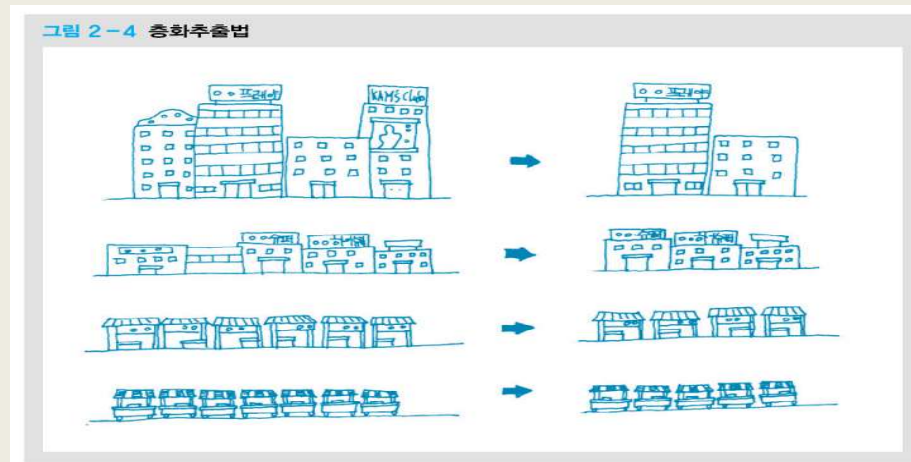
그림 2-3 집락추출법



# 표본추출방법

## (4) 층화추출법

- ❑ 모집단이 상당히 이질적인 원소들로 구성되어 있을 때 표본이 각 계층을 고루 대표할 수 있도록 표본을 추출하는 방법.
- ❑ 이질적인 모집단의 원소들을 서로 유사한 것끼리 몇 개의 층(stratum)으로 나눈 후에 각 층에서 표본을 랜덤하게 추출하는 방법.





❑ 예) 서울시내 슈퍼마켓의 연평균 매출액 조사 : 400개 표본

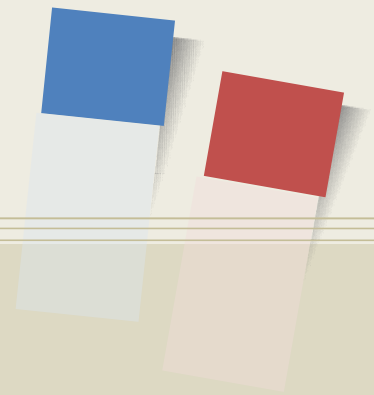
▷ 대형 50개	$\frac{50}{7550} \times 400 \approx 3\text{개}$
▷ 중형 500개	$\frac{500}{7550} \times 400 \approx 26\text{개}$
▷ 소형 1000개	$\frac{1000}{7550} \times 400 \approx 53\text{개}$
▷ 미니 6000개	$\frac{6000}{7550} \times 400 \approx 318\text{개}$


-----  
계 7550개






## 2.3절. 자료의 그래프적 표현





## 자료의 그래프적 표현



- ❑ 통계학에 대한 전문적인 지식이 없는 경우에도 쉽게 자료의 특성을 파악할 수 있도록 하는 기초적인 자료 요약방법이다.
- ❑ 대표적으로 도수분포표, 히스토그램, 막대그림표, 원그림표, 상자그림(box plot), 줄기와 잎 그림(stem-and-leaf plot) 등이 있다.

## 막대그래프와 원그림표

- 그림을 이용한 질적 자료의 표현방법은 막대그래프(bar chart)와 원그림표(pie chart)가 있다.
- **막대그래프**란 질적 자료에서 각 범주에 속한 관측도수를 기둥형태로 표현하는 방법으로 기둥의 크기에 의하여 상대적인 도수의 크기를 비교할 수 있다.
- **원그림표**는 각 범주의 관측도수의 상대적인 크기를 원을 분할한 형태로 표현하는 방법으로 파이(pie)를 분할하는 형태와 비슷하다고 하여 파이그림(pie chart)이라고 부르기도 한다.

# 원그림표의 작성방법

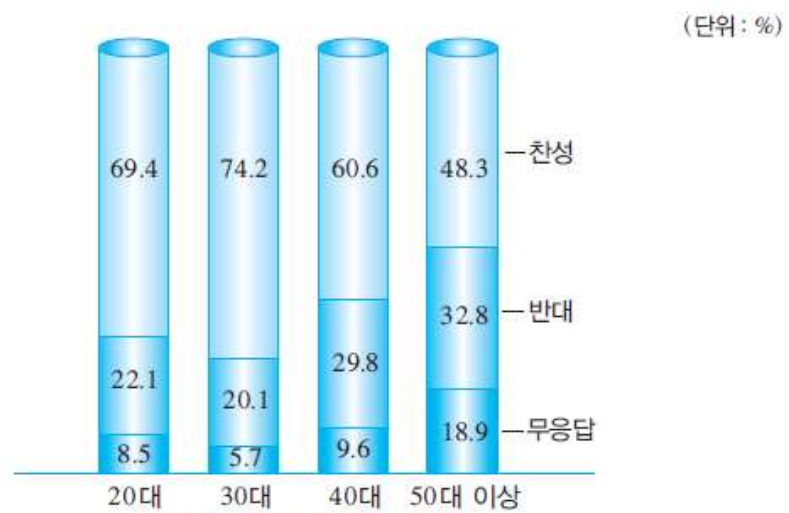
1. 적절한 수의 범주를 선택한다. 너무 많은 수의 범주를 선택하면 원그림표의 이해가 어렵게 된다.
2. 각 부분의 크기는 실제 관측도수의 비율과 같도록 그리고 크기 순서로 정리한다.



## 막대그림표의 작성방법

1. 막대의 폭은 모든 범주에서 동일하여야 한다. 만약에 어떤 범주는 폭을 크게 하고 다른 범주에서는 폭을 작게 하면 두 범주에 대한 객관적인 비교를 하는 데 장애가 된다.
2. 막대의 크기는 그 범주에 속한 관측도수의 크기에 비례하여야 한다. 예를 들면 10%가 관측된 범주의 기둥 크기는 20%가 관측된 범주의 막대크기의  $\frac{1}{2}$ 이어야 한다.
3. 막대의 모양은 자료를 의미할 수 있는 것으로 하는 것이 좋다. 예를 들면 자동차에 대한 조사를 할 때 자동차를 쌓아 놓은 모양으로 막대를 하면 이해에 도움이 된다.

그림 2-7 부유세 도입에 대한 찬반



자료 : 조선일보, 2004. 5. 10.

## 도수분포표와 히스토그램

- ❑ **도수분포표**는 숫자로 관측된 양적 자료(연속형 자료)를 일정한 구간으로 나눈 후에 각 구간에 속한 개수들의 수를 도수로 나타낸 표이다.
- ❑ **히스토그램**은 도수분포표에서 각 구간의 관측도수를 막대형태로 표현하여 그 크기를 비교할 수 있도록 하는 자료의 요약방법이다.

## 도수분포표와 히스토그램의 작성방법

1. 관측값 중에서 가장 작은 값과 가장 큰 값을 찾아내어 두 값 사이의 구간을 5~20개의 소 구간으로 나눈다. 단 이 소구간들은 다음 조건을 만족하여야 한다.
  - a. 각 관측값들은 하나의 소구간에만 속하여야 한다.
  - b. 구간의 경계선에는 관측값이 없어야 한다.
  - c. 소구간의 수에 대한 원칙은 없으나 관측값의 수에 따라 적절하게 선택한다.
2. 각 소구간에 속한 관측값의 수에 대한 상대도수를 계산한다. 여기에서 상대도수란 각 소구간의 관측도수를 전체 관측값의 수로 나눈 비율이다.

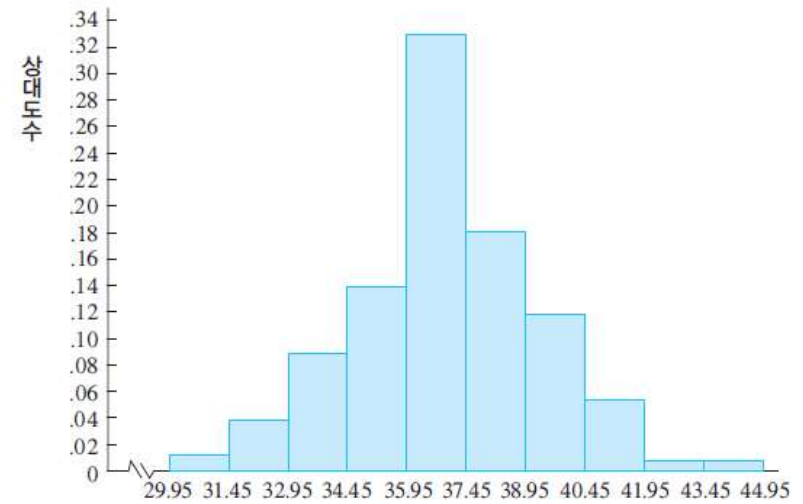


표 2-1 100개 차종의 연비 (단위 : mile/gal)									
36.3	41.0	36.9	37.1	44.9	36.8	30.0	37.2	42.1	36.7
32.7	37.3	41.2	36.6	32.9	36.5	33.2	37.4	37.5	33.6
40.5	36.5	37.6	33.9	40.2	36.4	37.7	37.7	40.0	34.2
36.2	37.9	36.0	37.9	35.9	38.2	38.3	35.7	35.6	35.1
38.5	39.0	35.5	34.8	38.6	39.4	35.3	34.4	38.8	39.7
36.3	36.8	32.5	36.4	40.5	36.6	36.1	38.2	38.4	39.3
41.0	31.8	37.3	33.1	37.0	37.6	37.0	38.7	39.0	35.8
37.0	37.2	40.7	37.4	37.1	37.8	35.9	35.6	36.7	34.5
37.1	40.3	36.7	37.0	33.9	40.1	38.0	35.2	34.8	39.5
39.9	36.9	32.9	33.8	39.8	34.0	36.8	35.0	38.1	36.9

표 2-2 100개 차종의 연비에 대한 도수분포표

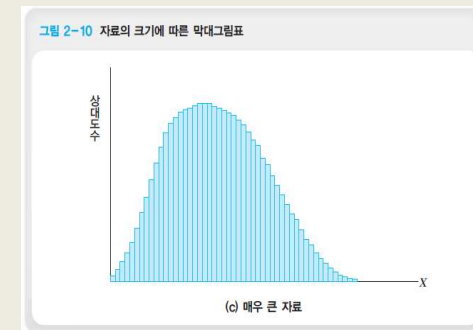
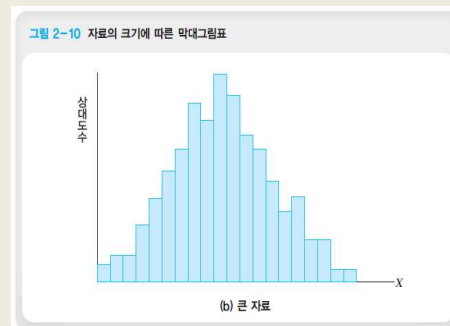
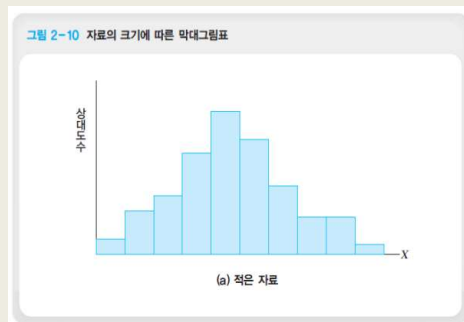
측정구간	도수	상대도수
29.95 ~ 31.45	1	.01
31.45 ~ 32.95	5	.05
32.95 ~ 34.45	9	.09
34.45 ~ 35.95	14	.14
35.95 ~ 37.45	33	.33
37.45 ~ 38.95	18	.18
38.95 ~ 40.45	12	.12
40.45 ~ 41.95	6	.06
41.95 ~ 43.45	1	.01
43.45 ~ 44.95	1	.01
총계	100	1.00

그림 2-9 100개 차종의 연비에 대한 막대그래프



# 자료의 크기에 따른 히스토그램

- 히스토그램은 관측값의 수가 증가함에 따라 소구간의 크기를 작게 함으로써 자료의 보다 정확한 표현을 가능하게 할 수 있다.
- 소구간의 크기가 매우 작은 경우에는 상대도수에 의한 히스토그램은 연속 곡선과 거의 유사하게 표현할 수 있는데, [그림 2-10]에 있는 세 개의 서로 다른 히스토그램을 비교해 보면 자료의 크기에 따른 히스토그램의 변화를 알 수 있다.



## 줄기와 잎 그림

- ❑ **줄기와 잎 그림(Stem-and-Leaf Plot)**은 숫자로 관측된 양적 자료를 정리하는 방법으로 막대그림표와 유사하나 막대그림표에서는 얻을 수 없는 정보인 자료의 최소값, 최대값 그리고 각 구간 내부에 있어서의 자료의 분포에 대한 정확한 정보를 제공해 준다.
- ❑ 막대그림표로 그렸을 때의 분포를 나타내는데 막대그림표가 각 구간 내 자료의 형태에 대한 정보를 제공하지 못하는 데 비하여 줄기와 잎 그림은 자료의 요약에 따른 정보의 손실이 전혀 없다.

## 줄기와 앞 그림 작성방법

1. 관측값의 숫자 단위(1단위, 10단위, 100단위, ...)를 이용하여 숫자를 두 부분으로 나누어 앞 부분은 줄기로, 그리고 뒷 부분은 앞으로 한다.
2. 줄기의 숫자를 작은 것부터 크기 순서에 따라 열(columnize)로 나열한다.
3. 각 관측값을 그 숫자가 속한 위치의 줄기에 맞추어 앞 부분을 기록한다.
4. 만약 각 줄기에 너무 많은 관측값이 주어진다면 각 줄기에 두 줄을 할당하여 첫 줄에는 앞의 0, 1, 2, 3, 4를 기록하고, 둘째 줄에는 앞의 5, 6, 7, 8, 9를 기록한다.
5. 각 줄기 내의 앞의 값들은 작은 것부터 크기 순서로 정리한다.

표 2-3

통계학 과목 수강생의 학기말 성적

65	62	73	85	65	46	36	49	81	76
60	44	43	72	21	33	83	46	64	49
12	74	91	78	60	48	24	62	54	97
69	31	89	96	96	97	86	88	85	61
95	54	85	89	51	77	81	72	47	35

그림 2-11 통계학 과목 수강생 성적의 줄기와 잎 그림

도수	줄기(stem)	잎(leaf)
1	1	2
2	2	1 4
4	3	1 3 5 6
8	4	3 4 6 6 7 8 9 9
3	5	1 4 4
9	6	0 0 1 2 2 4 5 5 9
7	7	2 2 3 4 6 7 8
10	8	1 1 3 5 5 5 6 8 9 9
6	9	1 5 6 6 7 7

끝~~❤❤