

2022학년도 2학기

경영 통계학

담당교수: 백수정



학습 목표

1. 데이터의 특성에 맞는 시각화 기법을 이해한다
2. 데이터 시각화(그림, 표, 히스토그램 및 기타 차트)를 수행하고 이해한다.

데이터 시각적으로 묘사하기

- 통계량을 계산하고 그래프를 그리기 전에 데이터를 눈으로 보고 그것이 어떻게 수집되었는가를 머릿속에 그려보는 것이 좋다.
- 데이터로 만들 수 있는 그래프 표현에는 여러 가지가 있다. 어떤 도표는 정량적(quantitative) 데이터에 적합하고 어떤 것은 범주형(categorical) 데이터에 적합하다.
- 수학을 사용하지 않고 시각적(도표 및 그래프)으로 데이터세트의 특징을 제공한다.
- 수학을 사용하여 숫자형(통계량 또는 표)으로 데이터세트의 특징을 제공한다.
- 이 장에서는 몇가지 기본적인 차트를 소개하고, 언제 이러한 차트를 사용하는지, 보다 효과적으로 만들기 위한 방법은 무엇인지, 그리고 현혹적인 그래프가 되지 않도록 하기 위해서는 어떻게 해야하는지 등에 대해 가이드라인을 제공한다.

예비적 평가

- 우선 단일변량 데이터(어떤 하나의 변수가 n 개의 관측치로 되어 있는 경우)부터 시작

특성	해석
측정	측정의 단위는? 정수형인가 연속형인가? 결측치가 있는가? 정확도나 표본추출 방법에 문제는 없는가?
중심	데이터 값들의 중심은 어디인가? 어떤 데이터 값이 전형적인 또는 중간값인가?
변동성	데이터의 퍼짐이 어느 정도인가? 데이터 값들이 얼마나 퍼져있는가? 특이한 값들이 있는가?
형태	데이터 값들의 분포가 대칭적인가? 비대칭적인가? 중심이 아주 뾰족한가? 평평한가? 봉우리가 두 개인가?

줄기-잎 그림

- 소규모 데이터세트를 간단히 시각화하는 방법이 줄기-잎 그림(stem-and-leaf plot)이다. 줄기-잎 그림은 탐색적 자료분석(EDA; exploratory data analysis)의 도구로서 데이터의 핵심 특징을 직관적인 방식으로 표현한다.
- 줄기-잎그림은 기본적으로 도수표시 형태이지만, 표시 마크를 쓰지않고 숫자를 사용한다
- 두 자리 또는 세 자리 정수 데이터인 경우 줄기(stem)는 십의 자리 수이고, 잎(leaf)은 일의 자리 수이다.

줄기-잎 그림

- 예를 들어 44개 P/E 비율의 줄기-잎 그림은 다음과 같다.

```
0 | 7 9
1 | 0 0 0 0 1 1 1 2 3 3 3 4 4 5 6 6 6 6 7 7 8 8 9 9
2 | 0 0 1 1 2 3 4 6 6 7 8
3 | 1 7 7 8
4 | 2
5 | 0 9
```

- 네번째 줄기의 데이터 값은 31, 37, 37, 38이다.
- 줄기의 간격은 동일하게 한다(줄기에 해당 데이터가 없는 경우라도)

줄기-잎 그림

- 줄기-잎은 중심경향을 보여줄 수 있으며(44개 P/E 비율 중 24개가 10~19 줄기에 속해 있다), 퍼져 있는 정도도 보여준다(범위가 7부터 59까지이다).
- 이 예에서 잎의 숫자들은 크기순으로 정렬되어 있지만 반드시 그렇게 해야하는 것은 아니다.
- 줄기-잎 그림은 줄기와 잎의 숫자를 합함으로써 원데이터를 복원할 수 있는 장점이 있다.

점그림

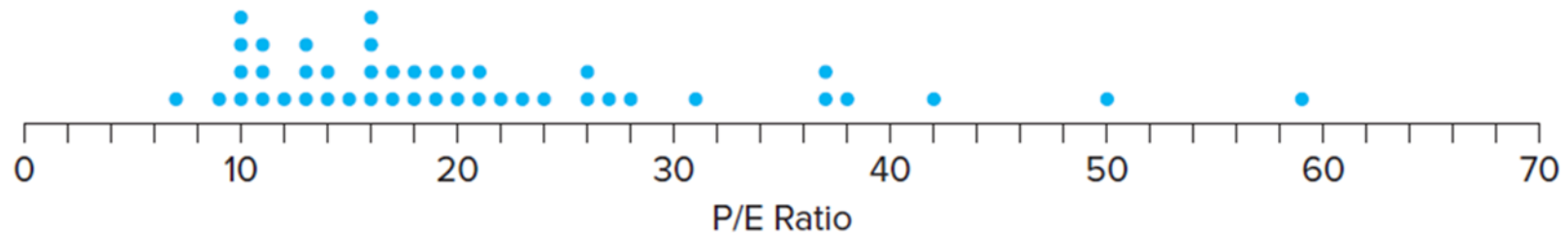
점그림(dot plot)은 숫자형 데이터를 가장 간단한 그래프로 나타내는 방법이다.

- 점그림은 이해하기 쉽다.
- 데이터의 퍼짐을 보여주고, 어디에 값들이 몰려 있는가를 나타냄으로써 대략적인 중심을 보여준다.

점그림

■ 점그림 만드는 기본 절차

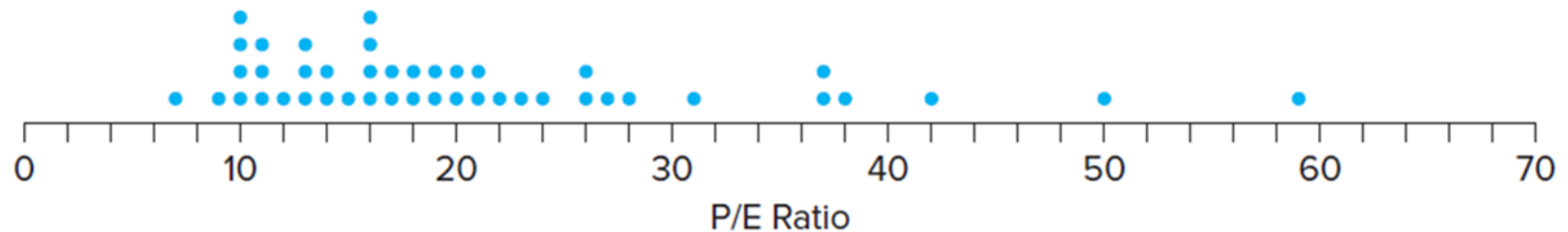
1. 데이터 범위를 포괄하는 척도를 만든다.
2. X 축을 만들어 척도를 표시한다.
3. 각 데이터 값을 척도 위에 점으로 표시한다.



- 만약 하나 이상의 값이 X 축 상에서 동일하거나 매우 비슷한 위치에 있게 되면 점을 세로로 쌓는 방식으로 그리게 된다.

점그림

- 범위가 7부터 59까지이다.
- 일부를 제외하고는 주로 10에서 25 사이에 있다.
- 대표적인 “중간” 데이터 값은 대략 17이나 18이다.
- P/E 비율이 큰 값들이 존재하여 데이터가 대칭적이지 않다



도수분포표

도수분포(빈도분포; frequency distribution)는 n 개의 데이터 값을 k 개의 계급(bin으로 분류한 표이다. 계급경계(bin limits)는 각 계급을 나누는 절사점(cutoff points)을 말한다. 계급경계에 따라 각 계급에 들어갈 값들이 결정된다. 계급의 폭은 모두 동일해야 한다.

- 기본 단계:
 1. 데이터를 오름차순으로 정렬한다.
 2. 계급(bin)의 수를 정한다.
 3. 계급경계를 정한다.
 4. 각 데이터의 값을 적정한 계급에 포함시킨다
 5. 표를 작성한다.

도수분포표

1단계: 데이터를 오름차순으로 정렬한다

- 최소값과 최대값을 찾는다

1단계: 데이터의 최솟값과 최댓값 찾기

정렬된 주가수익비율										
7	9	10	10	10	10	11	11	11	12	13
13	13	14	14	15	16	16	16	16	17	17
18	18	19	19	20	20	21	21	22	23	24
26	26	27	28	31	37	37	38	42	50	59

도수분포표

2단계: 계급의 수를 정한다.

- 계급의 수 k 는 표본의 수 n 보다 훨씬 작아야한다.
- 스터지스 룰(Sturges' Rule): 표본의 크기가 두배가 될 때 마다 계급수를 하나씩 늘린다.

표본크기(n)	바람직한 계급 개수(k)
16	5
32	6
64	7
128	8
256	9
512	10
1,024	11

도수분포표

3단계: 계급의 경계를 정한다.

- 데이터의 범위를 계급의 수로 나누어 계급의 적정 폭을 찾는다.

$$\text{계급 폭} \approx \frac{x_{\max} - x_{\min}}{k}$$

- 앞의 예에서 계급 폭을 계산해 보면

$$\text{계급 폭} \approx \frac{59 - 7}{6} = \frac{52}{6} = 8.67$$

- 여기에서 ‘깔끔한’ 경계값을 얻기 위해서는 위 숫자를 반올림하여 10으로 하고, 계급의 경계를 0, 10, 20, 30, 40, 50, 60으로 한다.

도수분포표

4단계: 각 계급내에 포함될 데이터 값을 센다.

- 일반적으로 아래쪽 경계는 계급에 포함시키고 위쪽 경계는 제외시킨다.
- 모든 계급이 중복되지 않아야 하고 각 데이터 값이 어느 한 계급에만 속하도록 해야 한다.

도수분포표

5단계: 표를 작성한다.

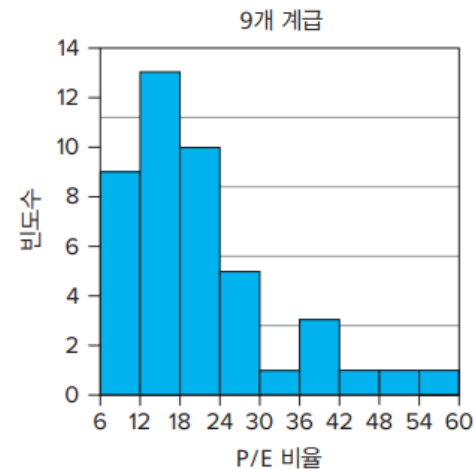
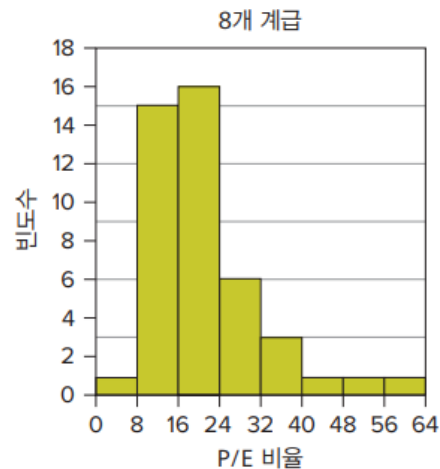
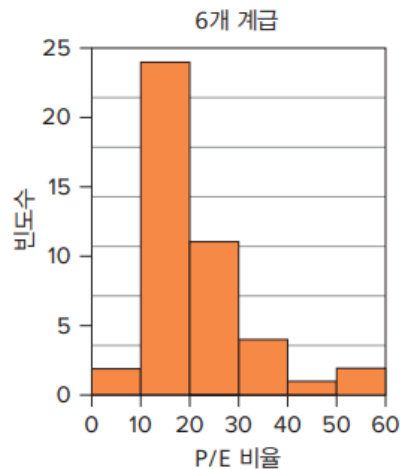
- 각 계급에 대해 절대적 도수만 제시할 수도 있고 상대도수나 누적(cumulative)도수를 포함시킬 수도 있다.

계급경계			상대			누적	
하한	상한	빈도(f)	빈도(f/n)	퍼센트	빈도	퍼센트	
0	< 10	2	$2/44 = 0.0455$	4.55	2	4.55	
10	< 20	24	$24/44 = 0.5455$	54.55	26	59.09	
20	< 30	11	$11/44 = 0.2500$	25.00	37	84.09	
30	< 40	4	$4/44 = 0.0909$	9.09	41	93.18	
40	< 50	1	$1/44 = 0.0227$	2.27	42	95.45	
50	< 60	2	$2/44 = 0.0455$	4.55	44	100.00	
		44		100.00			

히스토그램

- 히스토그램(histogram)은 도수분포를 그래프로 나타낸 것이다.
- 히스토그램은 막대도표(bar chart)이다.
- Y축은 각 계급의 도수수(또는 퍼센트)를 나타낸다.
- X축은 각 계급경계를 눈금으로 표시한다.

세 가지 종류의 P/E 비율 히스토그램 📁 PERatios



히스토그램

계급의 개수와 경계를 선택하는 일은 우리의 판단을 요구한다.
소프트웨어 프로그램으로 히스토그램을 그릴 수 있다. :

- Excel
- MegaStat
- Minitab

히스토그램

히스토그램을 통해 모집단의 형태를 짐작할 수 있다.

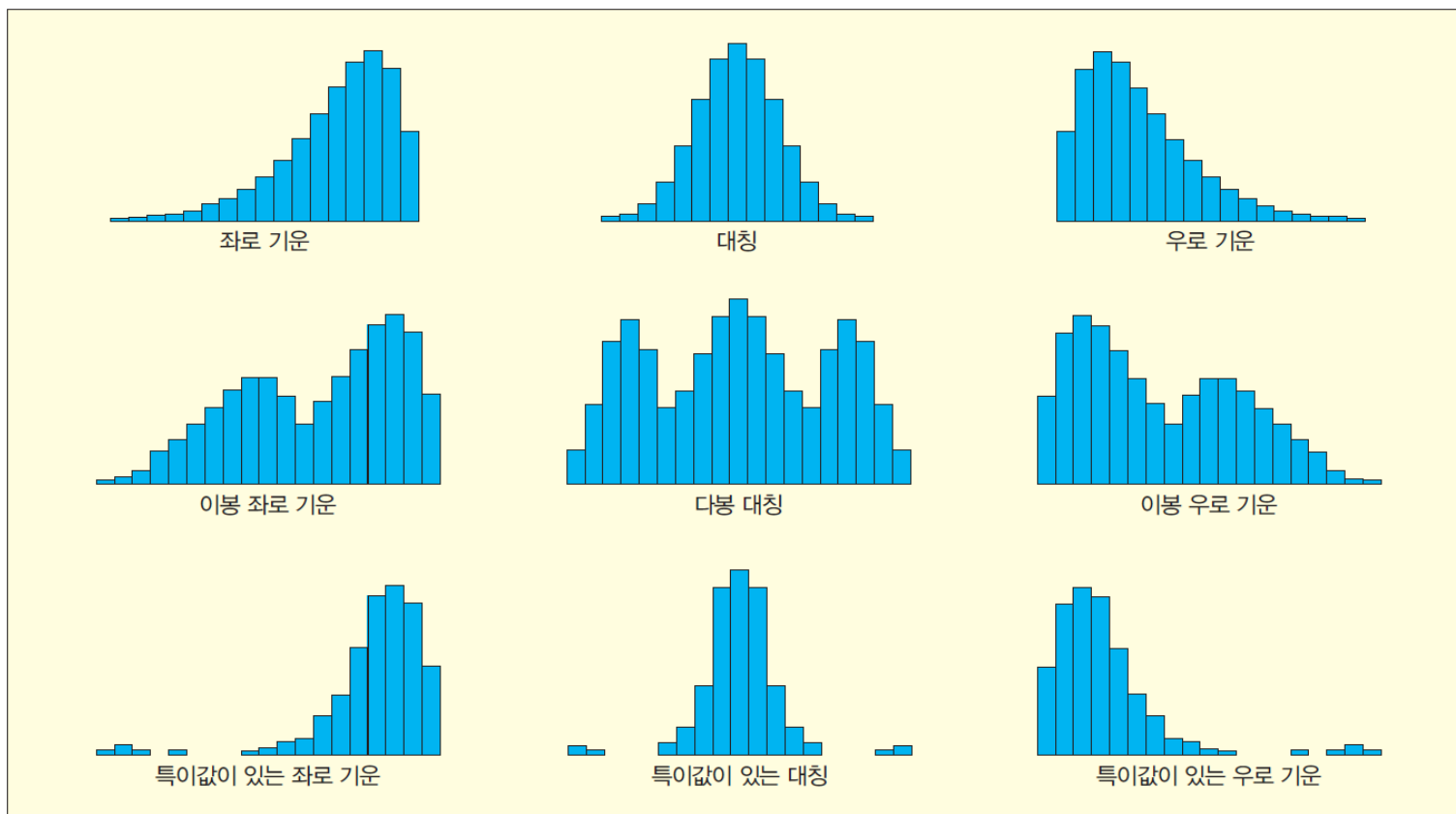
계급 개수나 경계 설정에 따라 모양이 달라질 수 있다.

왜도(skewness) – 히스토그램의 긴 꼬리가 어느 쪽에 있느냐에 따라 분류됨

- 왼쪽으로 기운(left-skewed) – (negatively skewed) 왼쪽꼬리가 긴 형태
- 오른쪽으로 기운(right-skewed) – (positively skewed) 오른쪽 꼬리가 긴 형태
- 대칭적(symmetric) – 양쪽 꼬리가 동일한 경우.

히스토그램

■ 히스토그램의 전형적인 모양



히스토그램

- 특이값(outlier)은 대다수의 데이터 무리에서 크게 벗어난 특이한 값을 말하는 것으로 특이한 원인에 의해 발생했거나 측정 오류 때문일 가능성이 높다.
- 특이값에 대해서는 다음 장에서 정식으로 논한다.
- 당분간은 특이값이란 히스토그램의 꼬리에 존재하는 비정상적인 점으로 이해하기로 하자.

효율적인 히스토그램 작성 요령

1. 계급의 수는 Sturges' Rule을 먼저 체크하되, 유동적으로 한다.
2. 적절한 계급의 폭을 정한다.
3. 계급 폭의 배수로 계급의 경계값을 정한다.
4. 모든 범위가 포함되었는지 확인하고 필요에 따라 계급의 수를 더한다.
5. 한쪽으로 기운 데이터의 경우에는 더 많은 계급을 필요로 할 수 있다.

도수 다각형 및 누적도수곡선

도수 다각형(frequency polygon)은 히스토그램 각 구간의 중간점, 그리고 히스토그램 구간의 시작과 끝을 0으로 하는 점들을 이은 선이다.

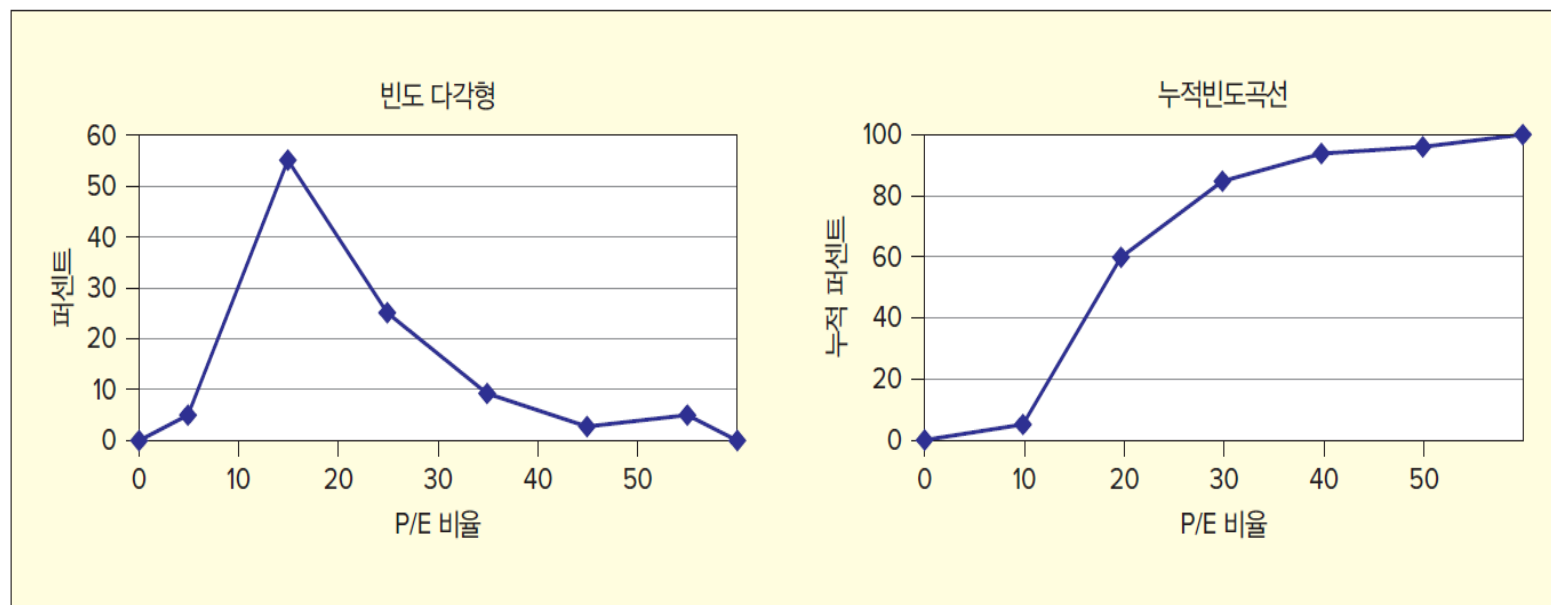
- 사실상 히스토그램과 동일하지만 두 개의 데이터세트를 비교하기에 유용하다(하나의 그림에 한 개 이상의 도수 다각형을 그릴 수 있기 때문).

누적도수곡선(ogive: pronounced “oh-jive”)은 누적도수를 선그림으로 나타낸 것이다.

- 백분위수를 파악하는 데 용이하고, 또는 표본의 형태를 정규분포(다음 장에서 소개)와 같은 알려진 분포와 비교하는 데 효과적이다.

도수 다각형 및 누적도수곡선

빈도 다각형 및 누적빈도곡선 📁 PERatios

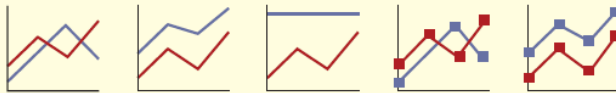


효과적인 엑셀 차트

- 엑셀은 강력한 그래프 기능을 갖추고 있다.
- 엑셀 차트들은 “삽입”에서 선택할 수 있다.

엑셀 차트 유형

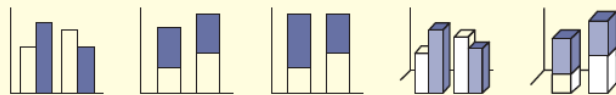
Line Charts



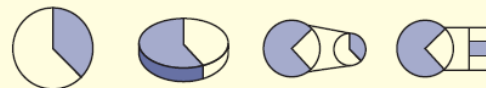
Scatter Plots



Column Charts



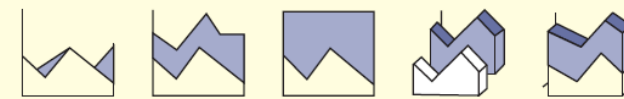
Pie Charts



Bar Charts

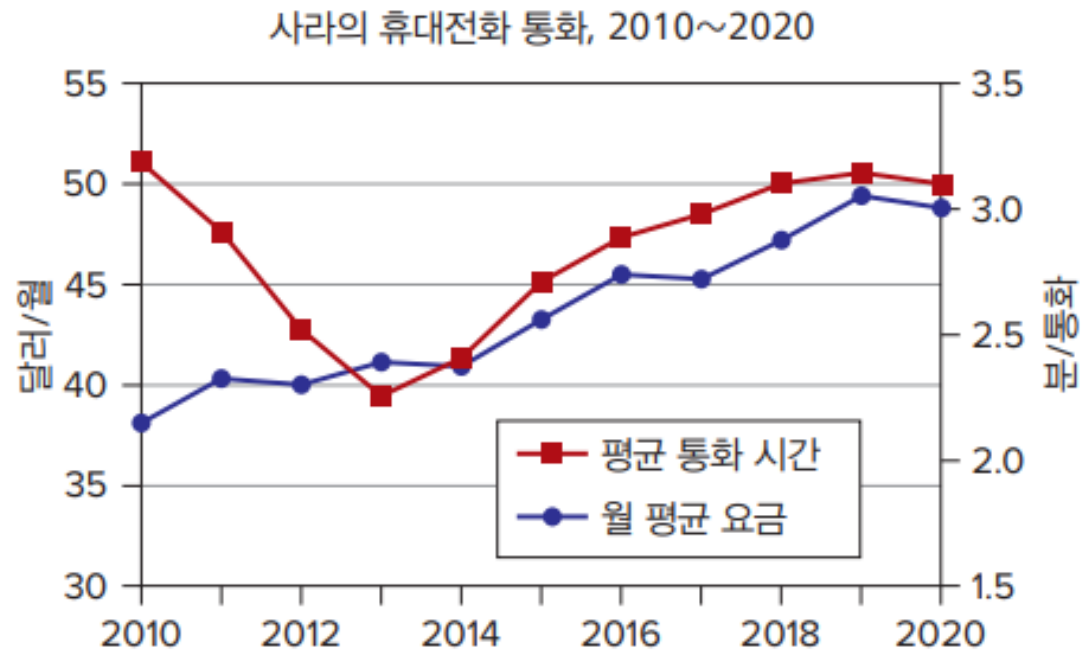


Area Charts



선도표(Line Chart)

- 시계열을 표시하거나, 트렌드를 나타내기 위해 사용된다.
- 선도표는 여러 변수를 하나의 그림에 나타낼 수 있다.



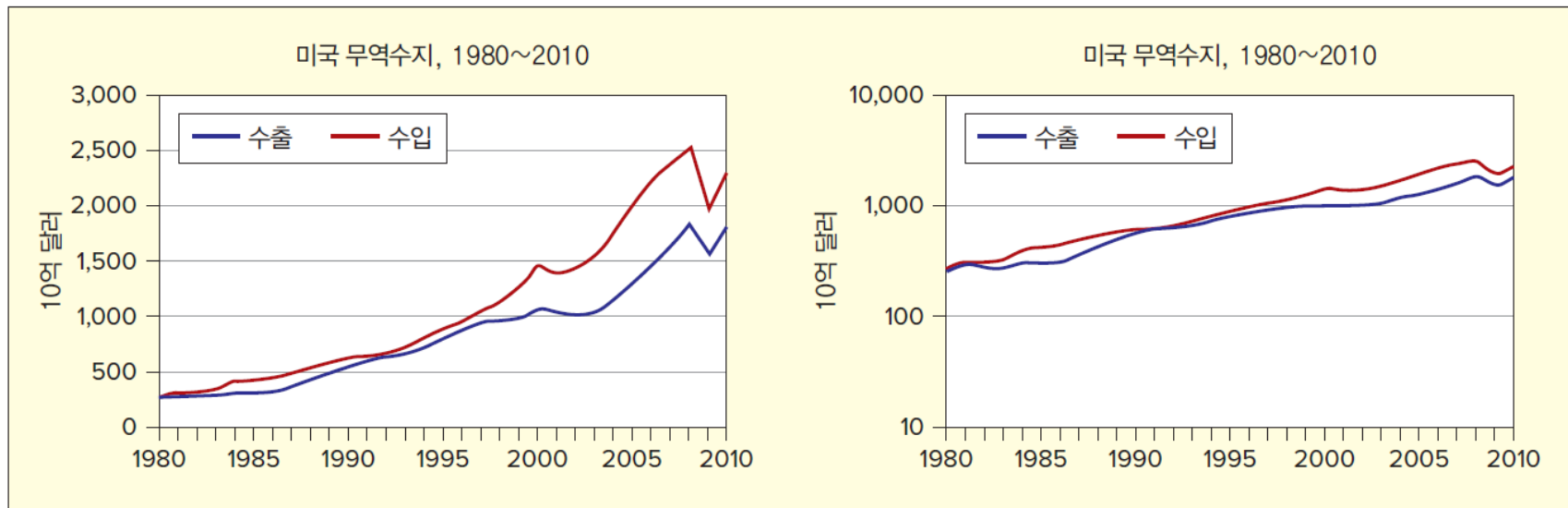
로그척도(Log Scales)

- 일반적으로 우리가 사용하는 척도인 산술척도(arithmetic scale)에서 Y축의 거리는 변수 값의 크기에 비례한다.
- 이와 달리 로그척도(logarithmic scale)에서는 동일한 거리가 동일한 비율을 나타낸다(이 때문에 로그척도를 비율척도라고도 부른다).
- 데이터의 값이 아주 크게 변할 때, 예컨대 10배 이상 변하는 상황에서 세로축에 로그척도를 쓰는 것이 좋다.
- 크기가 작은 값들을 보다 자세히 나타낼 수 있다.

로그척도(Log Scales)

- 로그척도는 증가가 누적적으로 지속되는 시계열 데이터에 유용하다(예: GDP, 국가부채, 자신의 소득)
- 로그 그래프는 어떤 변수의 증가 속도가 점차 빨라지는지(볼록 함수), 일정한지(직선), 점차 둔화되는지(오목 함수)를 보여준다.

산술척도와 로그척도 비교 📊 USTrade



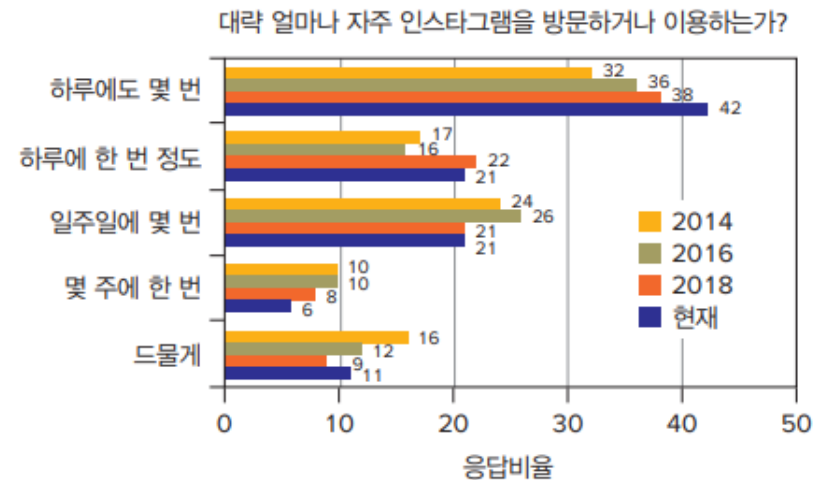
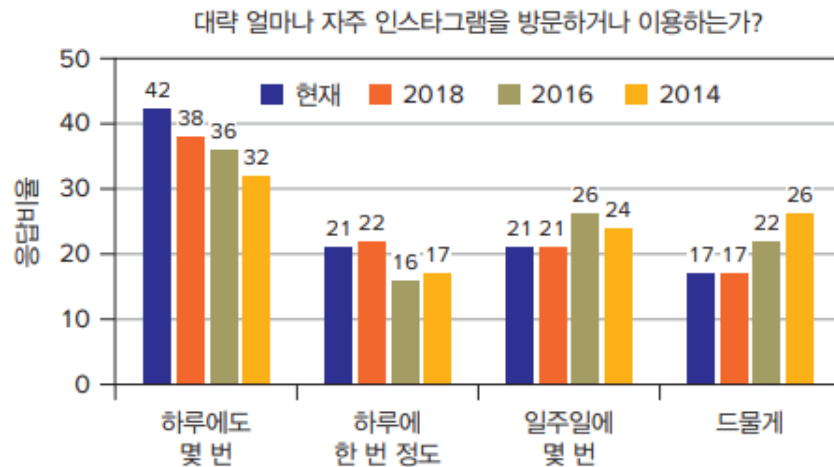
효율적 선도표 작성 요령

1. 선도표는 시계열자료에 사용 (횡단면자료에는 사용하지 않음)
2. 수치자료는 세로축에, 시간자료는 가로축에 왼쪽에서 오른쪽으로 표시. 기업에서는 대체로 이러한 패턴을 선호함.
3. 로그척도가 아니라면, 특별한 경우를 제외하고 세로축은 0에서 시작함(엑셀의 기본설정). 기업의 연례보고서와 주식투자 전망 등의 보고서에서는 이것이 의무화되어 있음.
4. 그래프 상의 혼잡을 피하기 위해 수치의 단위는 생략. 특히 시계열자료의 경우. 독자의 편의를 위해 중간선을 표시.
5. 데이터를 표시하는 마커(사각형, 삼각형 원 등)는 유용할 수 있으나, 자료의 수가 많거나 변수의 수가 많은 경우 그래프가 혼잡해 질 수 있음.
6. 선의 두께가 너무 두꺼우면 정확한 값을 확인하는데 어려울 수 있음.

막대도표

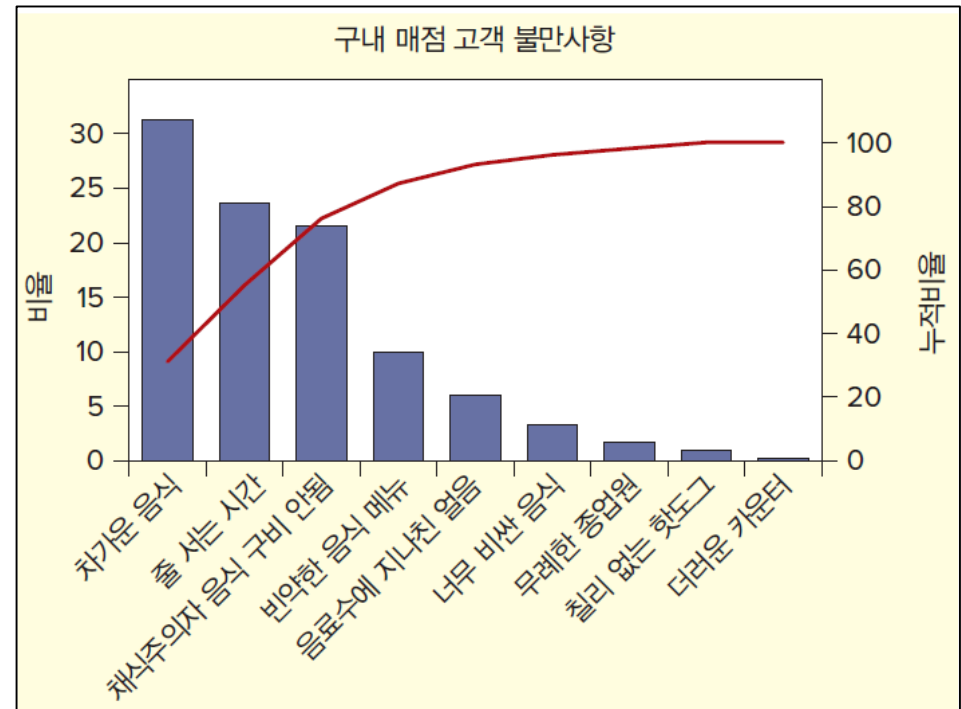
- 세로막대도표(column chart)는 데이터를 세로로 표시
- 가로막대도표(bar chart)는 데이터를 가로로 표시

동일 데이터를 두 가지 막대도표로 표현 📁 Instagram



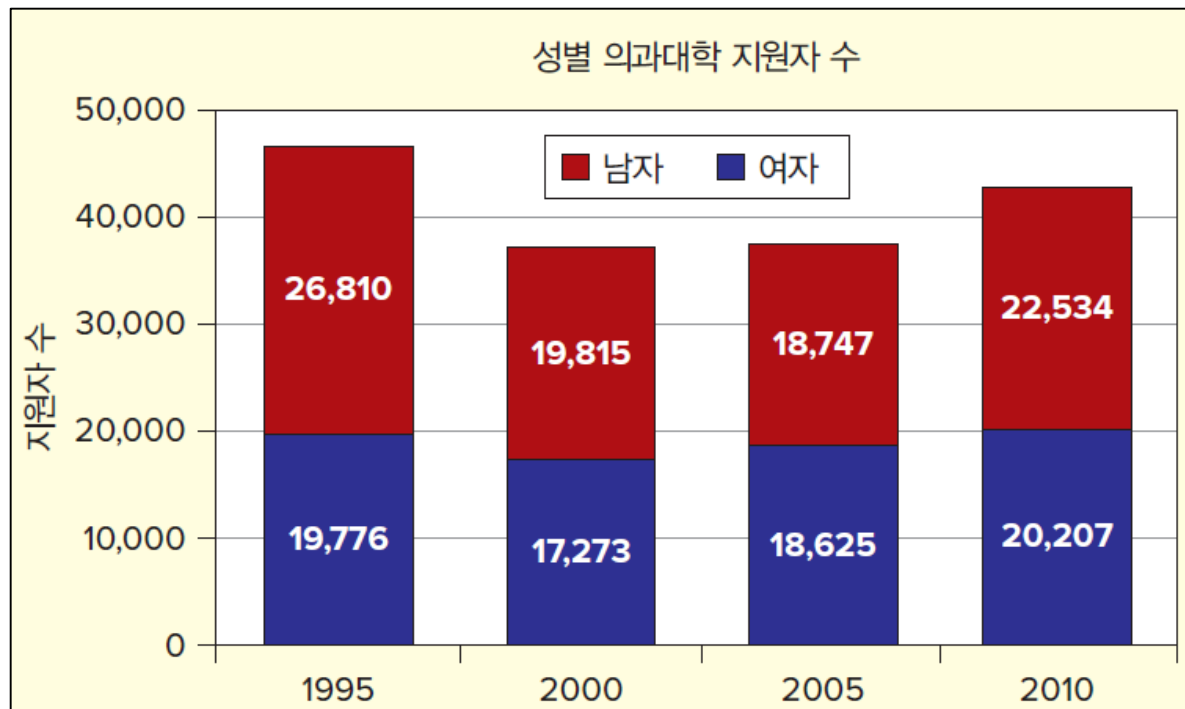
파레토 도표(Pareto Chart)

- 파레토 도표는 품질관리에서 결함 또는 오류의 도수를 나타내는 데 자주 이용된다.
- 범주는 도수에 따라 내림차순으로 정렬된다.
- 가장 빈번한 오류 또는 결함 등 중요한 몇 개 항목에만 집중하게 만들 수 있다



누적 막대도표(Stacked Column Chart)

- 막대의 최종 높이는 각 그룹별 크기를 모두 합한 것이다. 막대는 각 그룹에 색깔을 부여함으로써 그룹별 패턴을 보여줌과 동시에 막대 그 자체도 패턴을 보여준다.



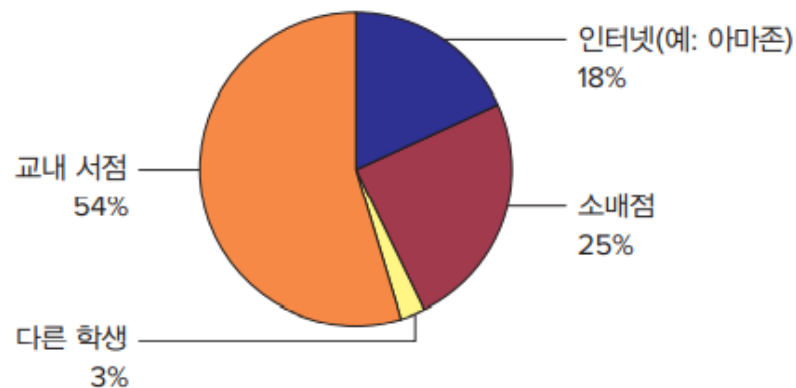
효율적 막대도표 작성 요령

1. 수치자료는 세로축에 범주자료는 가로축에 표시한다.
2. 범주자료가 시간자료라면 가로축에 왼쪽에서 오른쪽으로 표시.
3. 막대의 높이나 길이는 표시된 수치에 비례해야한다. 대부분의 통계패키지에서 기본설정이 되어 있기 때문에 어렵지 않다. 세로축을 0에서 시작하는 것은기업의 연례보고서와 주식투자 전망 등의 보고서에서는 의무화되어 있다.(예를 들어 수익을 과장되게 표현하는 것을 막기 위해) 그러나 보다 자세한 내용을 보여주어야 할 때는 예외일 수 있다.
4. 변수가 많거나 자료가 많아서 잘 보이지 않을 경우, 혹은 일반독자를 위해 단순한 자료가 필요한 경우가 아니라면 막대 위에 수치를 표시할 수 있다.

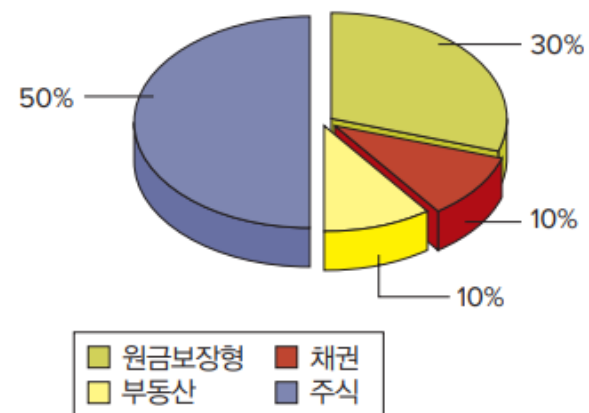
파이도표(Pie Chart)

- 파이도표는 데이터의 대략적인 아이디어만 전달할 수 있다.
- 파이도표를 올바르게 사용하기 위해서는 합이 전체가 되는 데이터이어야 한다(예: 시장점유율).
- 조각의 수는 적어야 한다(2-5개)
- 각 조각에 수치자료 혹은 비율을 표시한다.

자신의 통계학 교과서를 어디에서 구매했는가?



약간 보수적인 투자 포트폴리오




산포도

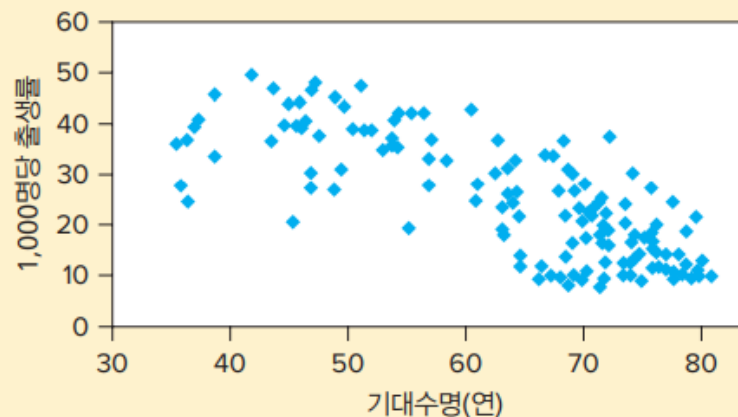
- 산포도는 표로는 분명하지 않은 한 쌍의 데이터 (X , Y)의 패턴을 시각적으로 보여준다.
- 산포도는 이변량 데이터 분석의 시작점이다.
- 우리는 두 변수의 관계를 살펴보기 위해 산포도를 이용한다.

산포도

- 그림 3.16은 X 축을 기대수명, Y 축을 출생률로 하여 산포도를 그린 것이다. 이 그림으로 볼 때 X 와 Y 사이에 어떤 연관성이 있어 보인다. 즉 출생률이 높은(낮은) 나라일수록 기대수명이 낮은(높은) 경향이 있는 것이다. 그러나 이것이 두 변수 간의 원인 - 결과를 의미하는 것은 아니다. 왜냐하면 두 변수 모두 제 3의 변수(예: 1인당 GDP)의 영향을 받을 수 있기 때문이다.

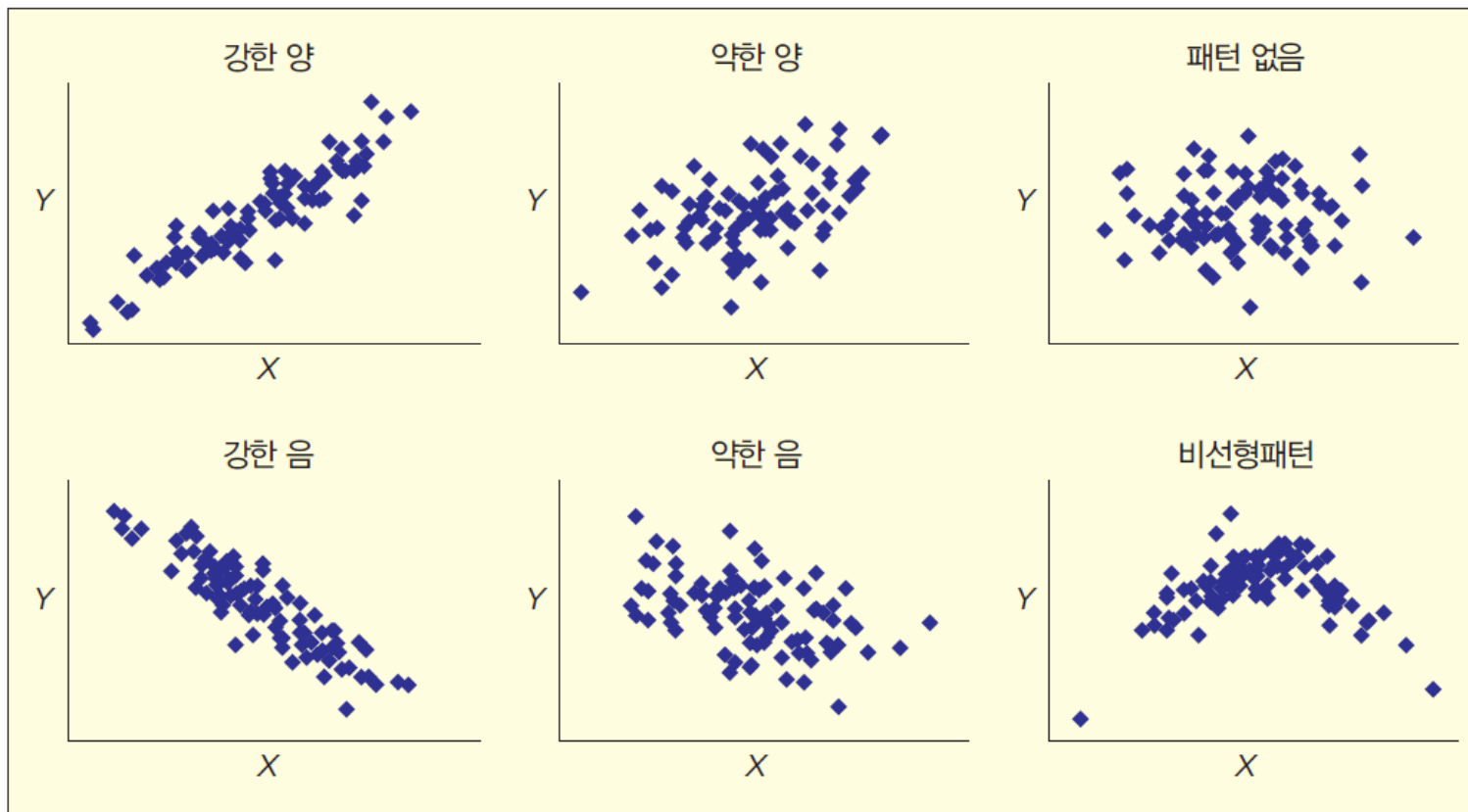
그림 3.16

출생률과 기대수명의 산포도($n = 153$ 개국)  BirthLife



산포도

- 아래 그림은 산포도의 몇 가지 전형적인 유형을 보여준다.



표

- 표(table)는 데이터를 표현하는 가장 간단한 형태이다.
- 가로 세로로 숫자를 잘 배열하여 표를 만들면 의미가 확실해지고 한 눈에 이해가 가능해진다.
- 의미를 명확히 전달하기 위해 데이터를 가로와 세로로 적절히 배치한다.
- 시계열 자료는 열을 따라 아래로, 비교하고자 하는 변수들은 행에 배치하면 집중도를 높일 수 있다.

표

- XY은행의 수익보고서 요약 (백만 달러)

	연도				
	2020	2019	2018	2017	2016
순이자 수익	2,306	2,025	1,805	1,772	1,652
대손충당금	2	77	237	138	28
비이자 수익	1,046	1,069	1,174	1,052	842
비이자 비용	1,821	1,748	1,915	1,815	1,586
소득세 충당금	307	518	178	244	292
순이익	1,185	735	462	530	586

효율적 표 작성 요령

1. 표는 단순하게, 그리고 목적을 잘 표현할 수 있도록 작성하라. 표가 크거나 많을 경우에는 요약표를 본문에 넣고 세부적인 표는 부록으로 처리하라.
2. 데이터를 행보다는 열끼리 비교할 수 있도록 배치하라.
3. 프레젠테이션 목적이라면 숫자들을 3~4 자리수가 되도록 반올림하라(예: 142.213보다는 142).
4. 자신이 강조하고 싶은 곳에 눈길이 가도록 표를 디자인하라.
5. 행이나 열의 제목은 간단하면서도 의미를 전달할 수 있어야 한다.
6. 동일한 열에서는 소수점 자리수가 동일해야 한다, 정렬은 소수점을 기준으로 한다.

피벗 테이블(Pivot Table)

- 피벗테이블은 엑셀에서 작성할 수 있다.
- 행과 열의 변수는 범주형이거나 이산수치형이어야 하고, 표의 내용은 수치가 되어야 한다.
- 일단 표가 작성되면 데이터 행렬로부터 변수이름을 드래깅하여 바꿀 수 있다.

피벗 테이블(Pivot Table)

The screenshot shows an Excel spreadsheet with a PivotTable titled "Pivot Table for Worksheet". The PivotTable Fields task pane on the right shows the following configuration:

- Filters:** None
- Columns:** Filing Type
- Rows:** Child Exempt...
- Values:** Count of Tax %

The PivotTable displays the following data:

Count of Tax %	Filing Type	Head of Household	Married Joint	Married Separate	Single	Grand Total
0		102	1150	59	1618	2929
1		308	425	10	21	764
2		196	555	4	9	764
3		43	228	1	2	274
4		16	41		1	58
5		1	7			8
6			3			3
10			1			1
Grand Total		666	2410	74	1651	4801

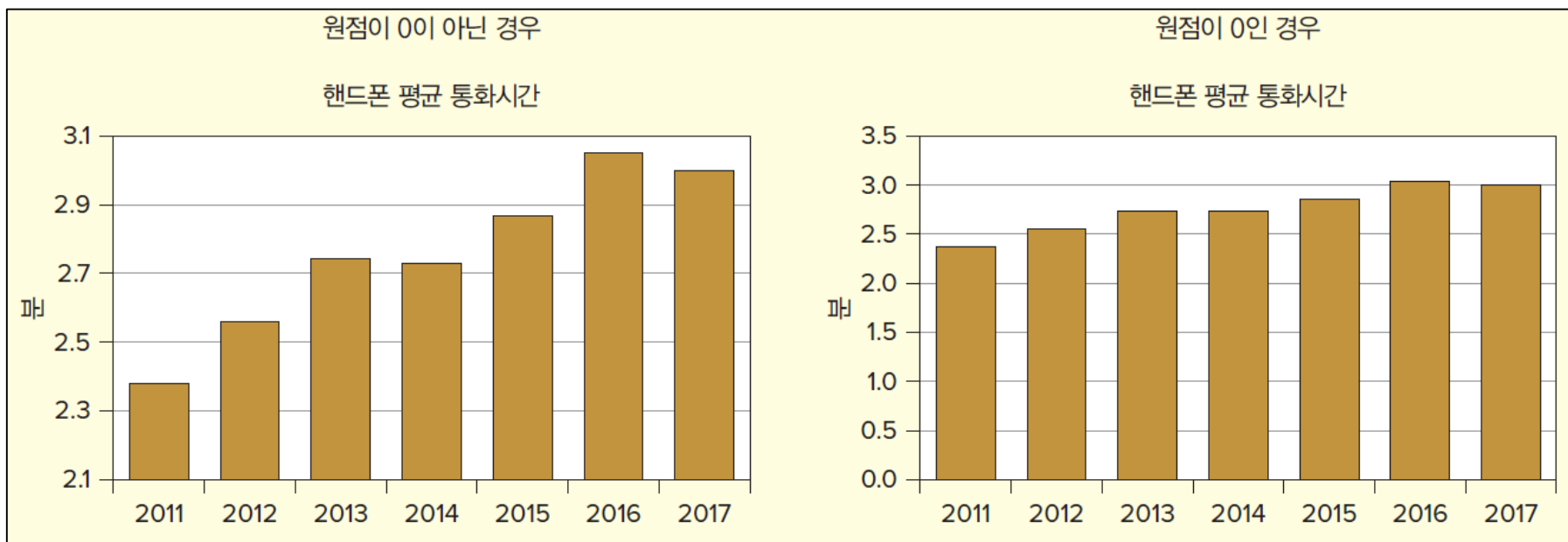
Below the main PivotTable, there is a summary table showing the average tax percentage for each filing type and exemption level:

Average of Tax %	Filing Type	Head of Household	Married Joint	Married Separate	Single	Grand Total
0		11.40	27.00	15.31	13.96	19.02
1		10.95	15.77	15.16	9.32	13.64
2		7.60	16.17	11.28	5.22	13.81
3		8.84	17.62	3.69	107.49	16.85
4		10.48	9.56		2.35	9.69
5		3.75	7.53			7.06
6			4.02			4.02
10			0.00			0.00
Grand Total		9.87	21.25	14.91	13.96	17.06

현혹적 그래프

오류 1: 원점이 0이 아닌 경우

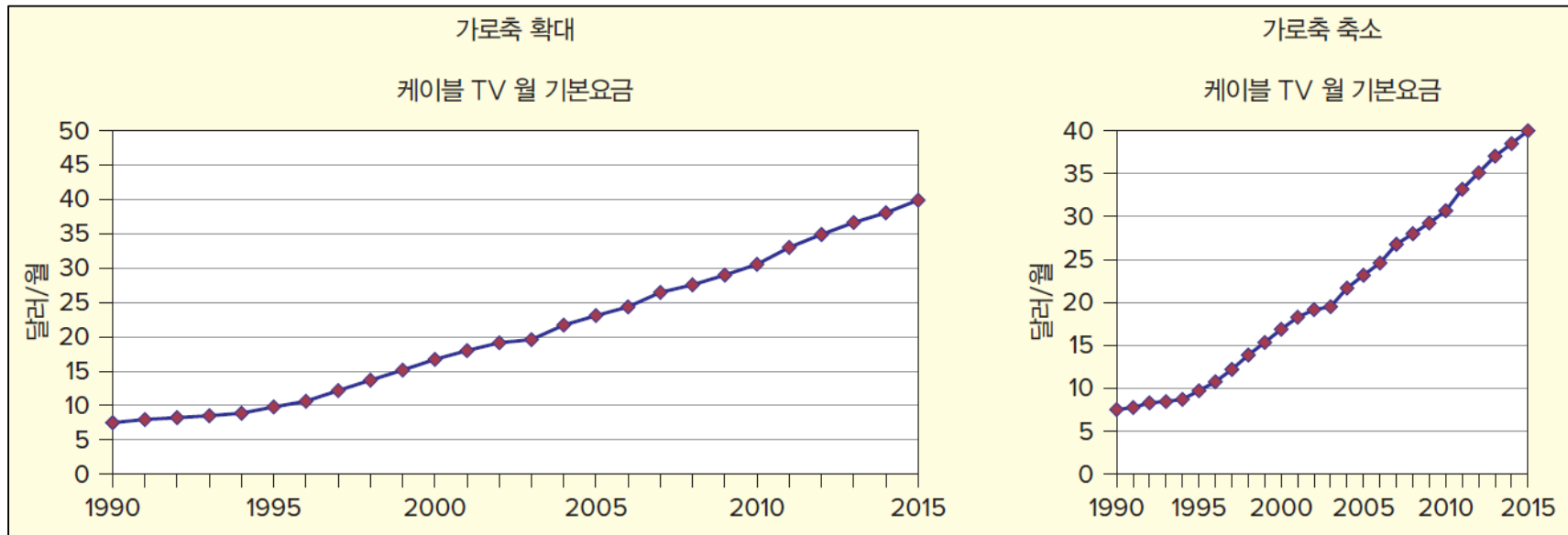
원점이 0이 아닐 경우 트렌드를 과장하게 된다.



현혹적 그래프

오류 2: 그래프 비율 조작

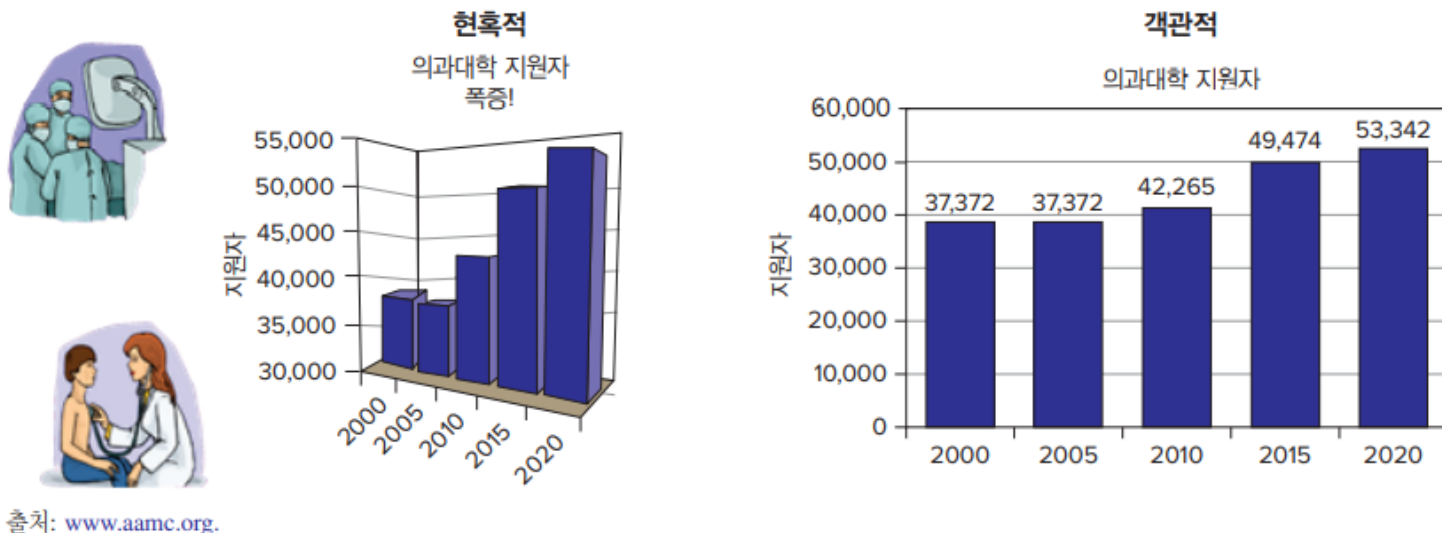
가로세로 비율(가로/세로)을 2.0 이하로 유지함
엑셀 그래프의 기본 가로세로 비율은 약 1.68임



현혹적 그래프

오류 3: 과장된 제목 및 산만한 그림

- 제목은 간결하면서도 목적에 적합해야 한다.
- 독자의 관심을 분산시키거나 감정을 자극하는 이미지를 피하라.
- 3-D차트는 시각적 효과는 좋으나 불분명해질 우려가 있다.



현혹적 그래프

오류 4: 단위 혹은 척도가 불분명한 경우

- 측정단위가 없거나 불분명한 경우 도표는 쓸모가 없어진다.
- 중간선은 독자들이 크기를 비교하는데는 도움이 되지만, 프레프를 복잡하게 하는 경우 생략되곤 한다.
- 막대도표에서 효율성을 극대화하기 위해서는 수치를 함께 표시해 준다.

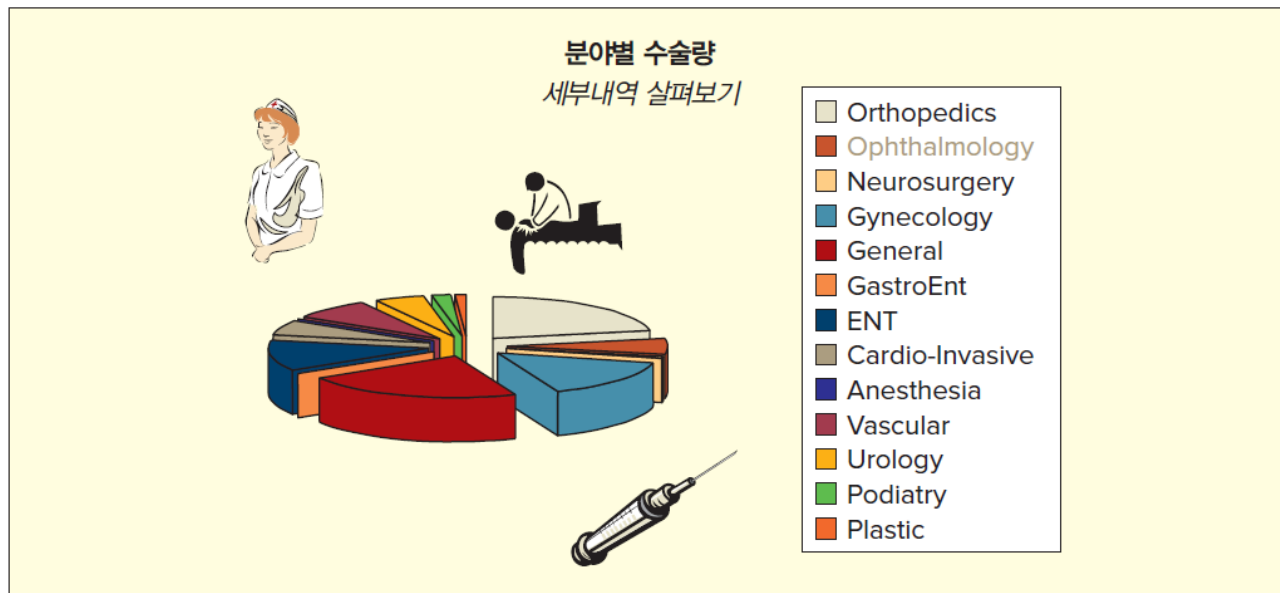
오류 5: 출처가 불분명한 경우

- 출처가 불분명한 경우 (예: “상무성”) 작성자가 출처를 모르거나 몇개의 출처를 혼용한 것으로 인식될 수 있다.
- 학술논문에서는 자료 인용 시 출처를 명확히 할 것을 요구한다.

현혹적 그래프

오류 6: 복잡한 그래프

- 복잡한 시각적 표현은 독자를 곤란하게 할 뿐이다.
- 무엇이 핵심 목표인지 명심하라.



출처: Hospital reports.

현혹적 그래프

오류 7: 불필요한 효과 넣기

- 슬라이드 쇼에서는 대개 주의를 모으기 위해 색채나 특수효과(소리, 화면전환, 텍스트 회전 등)를 이용한다.
- 그러나 일단 새로움이 사라지고 나면 사람들은 그와 같은 특수효과에 질리게 된다.

오류 8: 추정된 데이터

- ‘가장 최근의’ 숫자를 넣고 싶은 열정에서 시계열의 최근 데이터를 실적치가 아니라 추정하여 사용하는 경우가 있다. 이러한 경우 추정치라는 것을 밝히는 것이
- 최소한의 의무이다.

현혹적 그래프

오류 10: 면적 조작

- 막대의 높이가 높아짐에 따라 폭도 늘리게 되면 면적을 왜곡시킨다. 막대에 사람이라든지 동전 주유 펌프 등을 같이 그려 넣는 경우도 마찬가지다.

