기초 통계학

Bok, Jong Soon
javaexpert@nate.com
www.javaexpert.info

통계학이란 ?

- **統計學, Statistics**
- **수량적 비교를 기초로 하여, 많은 사실을 통계적으로 관찰하고 처리하는 방법을 연구하는 학문**
- **記述통계**
  - 측정이나 실험에서 수집한 자료의 정리, 표현, 요약, 해석 등을 통해 자료의 특성을 규명하는 통계적 방법
- **추리(추론)통계**
  - 기술통계로 어떤 모집단에서 구한 표본정보를 가지고 그 모집단의 특성 및 가능성 등을 추론해내는 통계적 방법

기술통계

- 관측을 통해 얻은 데이터에서 그 데이터의 특징을 뽑아내기 위한 기술
- 도수분포표나 히스토그램 등 표와 그래프로 표현하는 방법
- 평균값이나 표준편차 같은 통계량으로 표현하는 방법

추론(추리) 통계

- 통계학 방법과 확률 이론을 섞은 것
- 전체를 파악할 수 없을 정도의 큰 대상이나 아직 일어나지 않은 미래에 일어날 일에 관해 추측하는 것
- 부분으로 전체를 추측한다의 의미

도수분포표와 히스토그램

- 우리는 일상적으로 많든 적든, 의식하든 의식하지 못하든 데이터를 다루고 있다.
- 단순 데이터만으로는 아무것도 알 수 없다.
- 여대생 80명의 키를 정리한 데이터 → 여대생들의 키는 모두 같지 않고 제각각의 수치로 나타난다.
- **분포한다** → 다양한 수치로 나타나는 것

여대생 80명의 키.txt - Notepad

151	154	158	162
154	152	151	167
160	161	155	159
160	160	155	153
163	160	165	146
156	153	165	156
158	155	154	160
156	163	148	151
154	160	169	151
160	159	158	157
154	164	146	151
162	158	166	156
156	150	161	166
162	161	156	156
157	157	156	157
162	161	156	156
162	168	149	159
169	162	162	156
150	153	159	156

도수분포표와 히스토그램 (Cont.)

- 분포가 생기는 이유는 그 수치들이 결정된 이면에 어떤 **불확실성**이 움직이고 있기 때문이다.
- 불확실성의 구조가 제각각인 수치를 발생시킨다고 생각하는 것이다.
- 하지만,
- **불확실성**이라는 말로 표현하기는 하지만, 여기에도 고유한 **특징**이나 **반복되는 것**이 있다.
- **분포의 특성** → 그 고유한 특징이나 반복되는 것

도수분포표와 히스토그램 (Cont.)

- 결국 통계라는 것은 데이터 그 자체, 즉 현실 그 자체로부터 무엇인가 그 분포의 특징이나 반복되는 것을 이끌어 내기 위한 방법이다.
- 축약 : 데이터로 나열되어 있는 많은 숫자를 어떤 기준으로 정리정돈해서 의미있는 정보만을 추출하는 것
  - 그래프로 만들어서 그 특징을 파악할 수 있도록 한다.
  - 숫자 하나로 특징을 대표하도록 한다.
  - 여기서 대표하는 숫자를 통계량이라고 한다.

도수분포표와 히스토그램 (Cont.)

- **도수분포표 만들기**

1. 데이터 중에서 최대값과 최소값을 찾는다.
2. 최대값에서 최소값까지 포함되도록 구간을 자르기 좋은 대강의 범위를 만들고, 그 범위 내에서 5 ~ 8개 정도의 작은 범위(작은 구간)들로 자른다. 이렇게 자른 작은 범위를 계급이라고 한다.
3. 각 계급을 대표하는 수치를 정한다. 일반적으로 가장 가운데 값을 선택하는 경우가 많다. 이것을 계급값이라고 한다.

도수분포표와 히스토그램 (Cont.)

- **도수분포표 만들기**

4. 각 계급에 들어가 있는 데이터의 총 개수를 센다. 이것을 도수라고 한다.
5. 각 계급의 도수가 전체에서 차지하는 비율을 계산한다. 이것을 상대도수라고 한다. 상대도수는 합하면 1이 된다.
6. 어느 계급까지의 도수를 모두 합한다. 이것을 누적도수라고 한다. 최종 누적도수는 데이터의 총 개수와 일치한다.

도수분포표와 히스토그램 (Cont.)

- **도수분포표 만들기 – 여대생 80명의 키 데이터**

1. 최대값은 169, 최소값은 143이다.
2. 범위를 143과 가까운 구간에서 자르기 좋은 숫자로 140을 선택하고, 169와 가까운 구간에서 자르기 좋은 숫자로 170을 선택해서 140에서 170까지를 범위로 하는 계급을 만든다. 그리고 5개 데이터씩(5cm씩) 묶으면 6개의 계급이 생긴다.
3. 계급값으로는 가장 가운데 값을 사용한다. 예를 들면 141,142,143,144,145의 5개 중에서 가운데 값인 143을 선택한다. 이와 같이 모든 계급에서 대표값을 선택한다.

도수분포표와 히스토그램 (Cont.)

- **도수분포표 만들기 – 여대생 80명의 키 데이터**

4. 각 계급에 들어가 있는 데이터의 총 개수(도수)를 센다.
5. 각 도수를 데이터의 총 개수 80으로 나누어서 상대도수를 구한다.
6. 도수를 위에서부터 차례로 더해 내려가며 누적도수를 계산한다.

도수분포표와 히스토그램 (Cont.)

- **도수분포표 만들기 – 여대생 80명의 키 데이터**
  - 완성된 도수분포표

계급	계급값	도수	상대도수	누적도수
141 – 145	143	1	0.0125	1
146 – 150	148	6	0.075	7
150 – 155	153	19	0.2375	26
156 – 160	158	30	0.375	56
161 – 165	163	18	0.225	74
166 – 170	168	6	0.075	80

도수분포표와 히스토그램 (Cont.)

- 이렇게 도수분포표를 만들면 잃어버리는 정보가 있다.
- 어떤 정보를 잃었는가 → 데이터에 나타나 있던 수치들 자체
- 각 칸에는 도수만 나타나 있다. 그 세부적인 수치를 잃어버렸다.
- 그 이유는 ?
- 바로 도수분포표를 만들면서 생기는 축약으로 인해 발생한 일이다.
- 하지만, 이 축약된 수치로 데이터의 특징을 발견할 수 있다.

도수분포표와 히스토그램 (Cont.)

- **특징 1**
  - 키(데이터)는 균등하게(모두 똑같이) 분포하지 않고, 어느 한 곳에 (구체적으로 156 ~ 160의 계급에) 집중되어 있다.
- **특징 2**
  - 또한 집중되어 있는 곳을 기점으로 삼으면, 이 기점으로부터 작은 편에 속하든지 큰 편에 속하는 추이를 보인다. 즉, 데이터의 분포에는 어느 한 곳을 축으로 좌우 대칭성이 있다는 말이다.

도수분포표와 히스토그램 (Cont.)

- 여대생 80명의 키 데이터에서는 보이지 않았던 특징을 발견할 수 있었다.
- 축약은 결국, 데이터의 세부적인 수치들을 희생시키고 있지만, 이 희생으로 **데이터의 분포와 그 이면에 있는 특징들이 돋보이게 됐다.**
- 다른 말로 설명하자면...
- 데이터를 통해 요점을 찾았다고 볼 수 있다.
- 데이터의 요점 → 축약

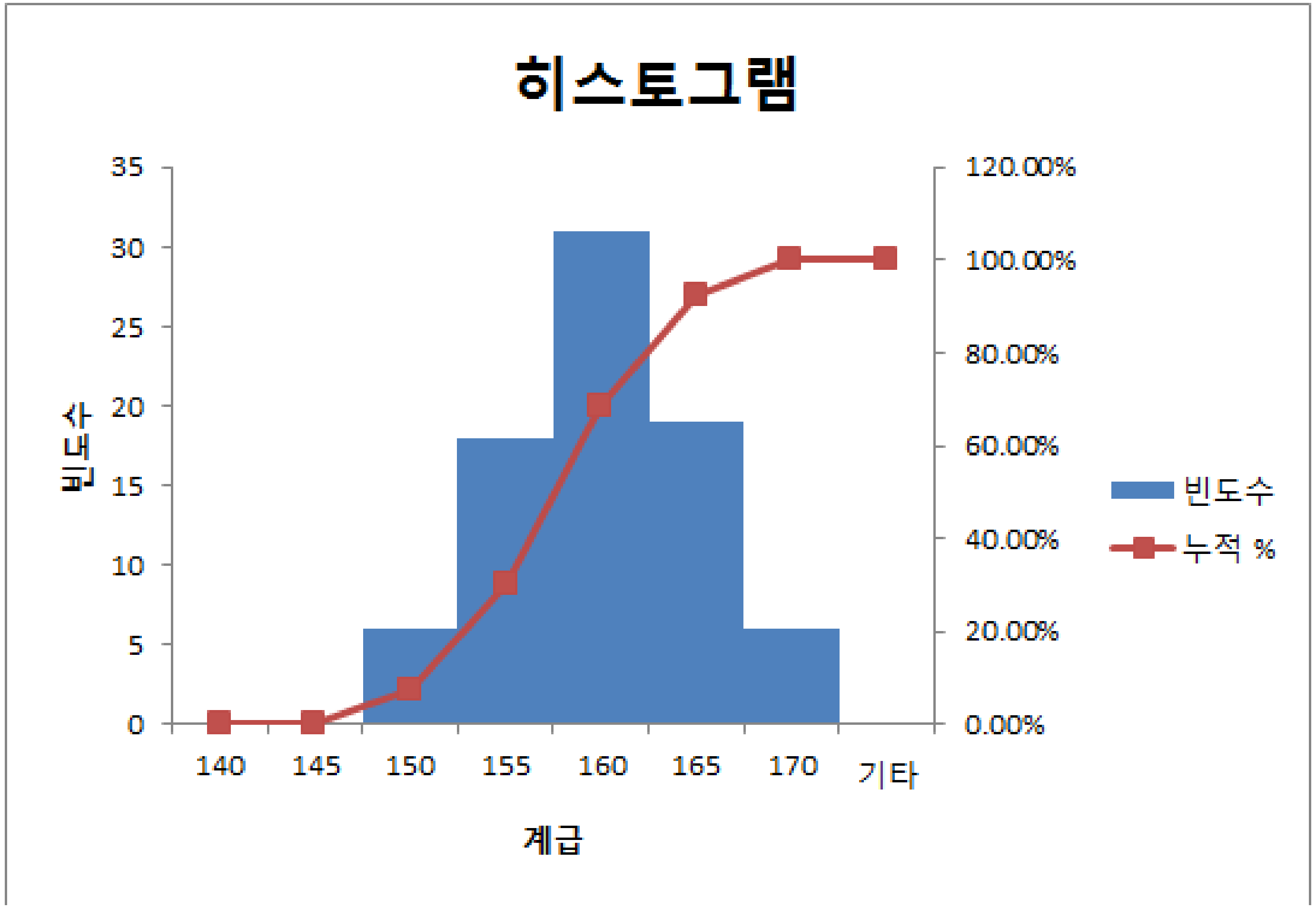
도수분포표와 히스토그램 (Cont.)

- **히스토그램 만들기**

  1. 가로축에 계급값(도수분포표 둘째 칸에 있는 수)을 같은 간격으로 둔다.
  2. 각 계급값 위에 막대를 세우는데, 막대 높이는 그 계급값에 속한 계급도수(도수분포표의 셋째 칸)로 한다.

도수분포표와 히스토그램 (Cont.)

- 히스토그램 만들기



연습문제

- 다음은 어느 학교의 몸무게 데이터이다. 도수분포표와 히스토그램을 작성하시오.

여대생 몸무게.txt - Notepad

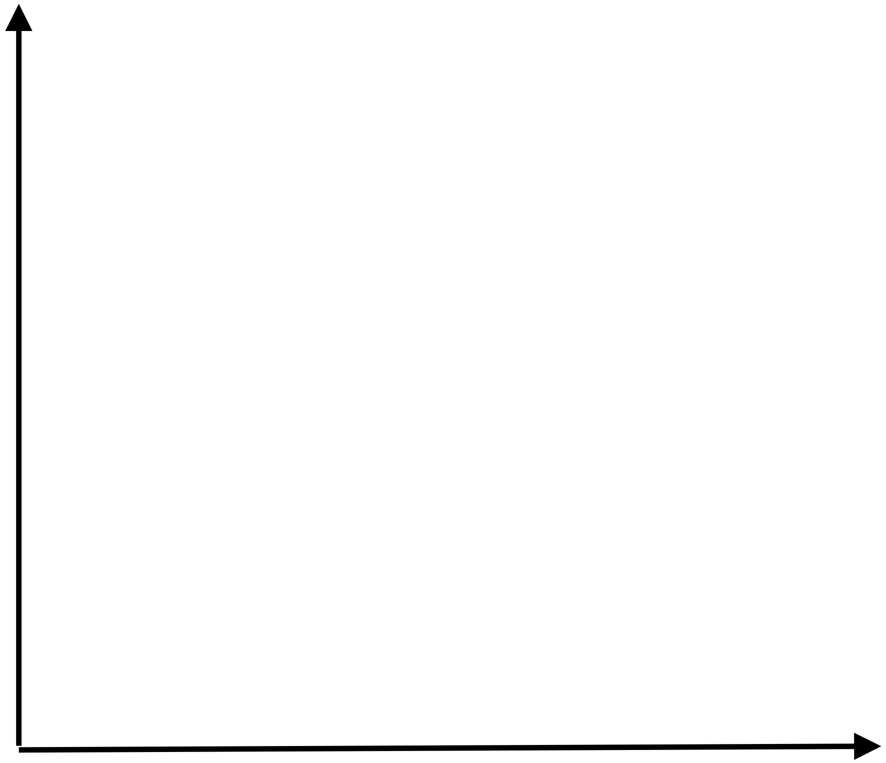
File Edit Format View Help

48	54	47	50	53	43	45	43
44	47	58	46	46	63	49	50
48	43	46	45	50	53	51	58
52	53	47	49	45	42	51	49
58	54	45	53	50	69	44	50
58	64	40	57	51	69	58	47
62	47	40	60	48	47	53	47
52	61	55	55	48	48	46	52
45	38	62	47	55	50	46	47
55	48	50	50	54	55	48	50

연습문제

- 다음은 어느 학교의 몸무게 데이터이다. 도수분포표와 히스토그램을 작성하시오.

계급	계급값	도수	상대도수	누적도수
36 – 40				
41 – 45				
46 – 50				
51 – 55				
56 – 60				
61 – 65				
66 – 70				



정리

- 데이터 자체는 현실 그대로를 나타내지만, 이것을 아무리 자세히 본다고 해도 알 수 있는 것은 없다.
- 데이터를 축약하는 방법에는 **그래프를 만드는 방법**과 **통계량을 구하는 방법** 2가지가 있다.
- 도수분포표는 데이터를 5 ~ 8개 정도의 그룹으로 나누는 것이다. **도수분포표로 데이터의 특성(데이터가 집중되는 곳이나 대칭성 등)을 파악할 수 있다.**
- 히스토그램이란 **도수분포표를 그래프로 바꾼 것**으로, 더욱 쉽게 데이터의 특징을 파악할 수 있다.

평균값의 역할과 평균값을 이해하는 방법

- **통계량은 데이터를 요약한 수치**
  - 도수분포표나 히스토그램의 단점
    - 그래프를 보고 데이터의 특징을 생각할 때 사람에 따라서 받아들이는 인상이 각각 다르다.
    - 도수분포표와 히스토그램은 상당히 많은 공간을 필요로 한다.
  - 위의 단점을 극복하기 위한 또 하나의 축약
  - 이것이 통계량이다.
- **통계량은 데이터의 특징을 하나의 숫자로 요약한 것**
- **즉, 데이터의 어떤 비슷한 특징을 요약하고 싶은가**

평균값의 역할과 평균값을 이해하는 방법 (Cont.)

- **평균값이란?**
  - 데이터 합계를 데이터 총 개수로 나누기해서 얻은 값
  - 예
    - {151 + 154 + … + 161} ÷ 80 = 157.575

평균값의 역할과 평균값을 이해하는 방법 (Cont.)

- **도수분포표에서 평균값**
  - (계급값 X 상대도수)를 계산해 합계를 구하면 평균값이 나온다.

A(계급값)	B(상대도수)	A X B
143	0.0125	1.7875
148	0.075	11.1
153	0.2375	36.3375
158	0.375	59.25
163	0.225	36.675
168	0.075	12.6

(A X B)의합계(평균값) : 157.75

평균값의 역할과 평균값을 이해하는 방법 (Cont.)

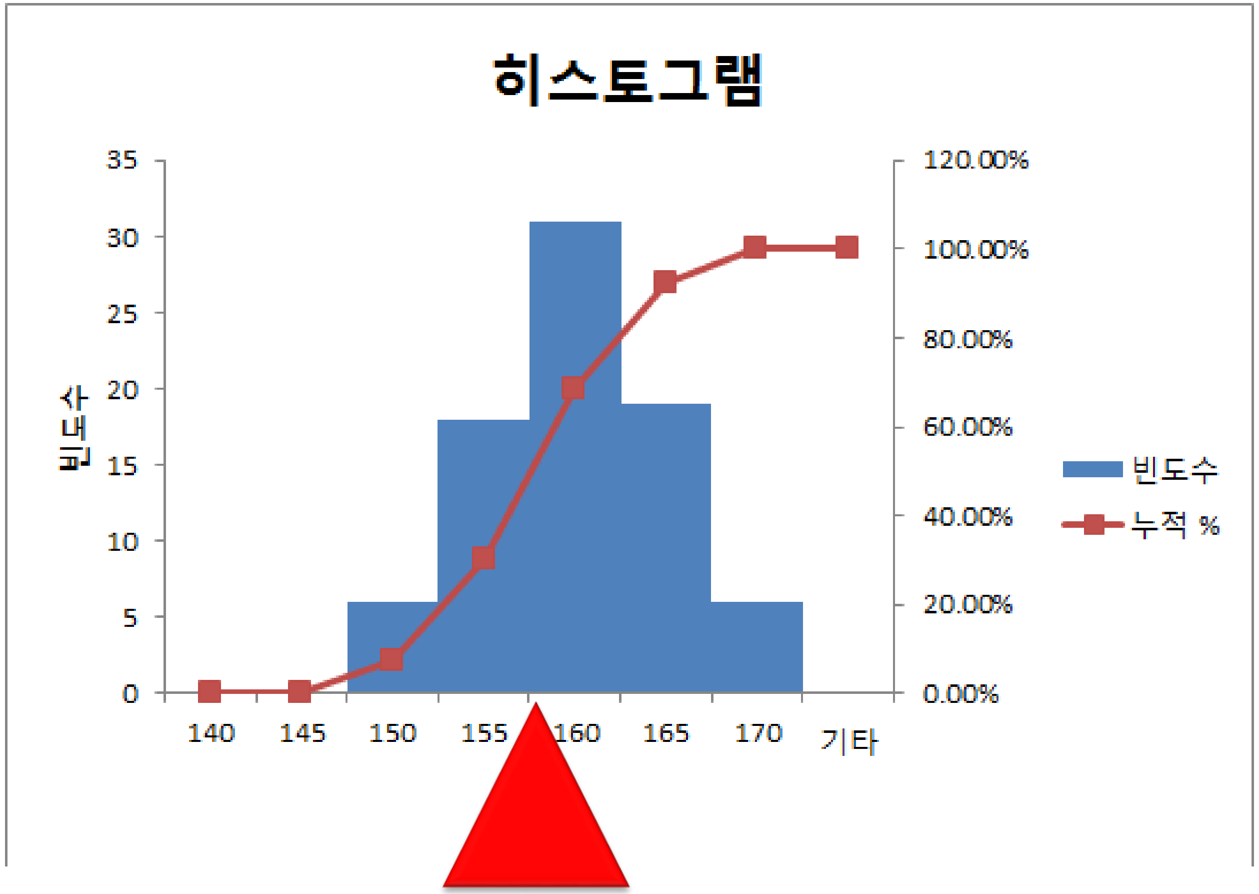
- **도수분포표에서 평균값**
  - 두 계산의 차이는 거의 나지 않는다.
  - 산술평균 : 157.575
  - 도수분포표에서의 계산 : 157.75
  - 도수분포표를 만드는 것이 평균값이라는 통계량에는 별로 큰 영향을 주지 않는다.
  - {(계급값 x 상대도수)의 합계}는 통계학 전반에 걸쳐 사용하는 것이기 때문에 꼭 기억할 것

평균값의 역할과 평균값을 이해하는 방법 (Cont.)

- **도수분포표에서 평균값**
  - 148 x 6 ÷ 80 = 148 x (6/80) = (계급값 x 상대도수)
  - 평균값 = {(계급값 x 도수)의 총합} ÷ (총 데이터 수)
    = (계급값) x {(도수 ÷ 총 데이터 수)의 총합}

평균값의 역할과 평균값을 이해하는 방법 (Cont.)

- **히스토그램에서 평균값의 역할**
  - 히스토그램을 지렛대라고 가정하면 평균값은 지렛대가 일자로 균형을 이루는 지점이다.
  - 예, 여대생 키의 평균값인 157.75위치에 삼각형 모양의 받침을 두면 좌우 균형을 이루게 된다는 뜻.



평균값의 역할과 평균값을 이해하는 방법 (Cont.)

- **평균값을 어떻게 이해해야 하는가?**
  - 데이터는 수치적으로 널리 퍼져있지만, 그 널리 퍼져있는 것 중에 하나의 수를 모든 데이터를 대표하는 수로 뽑은 것.
  - 데이터들은 평균값 주변에 분포되어 있다.
  - 많이 나타나는 데이터는 평균값에 주는 영향력이 크다
  - 히스토그램이 좌우대칭일 경우, 평균값은 대칭이 되는 축에 자리한다.
  - 평균값은 합계의 의미에서는 원래의 데이터로 보기에도 손색이 없을 정도의 수

연습문제

- 다음의 데이터로 도수분포표를 채우고 평균값을 구하시오.

계급값	도수	상대도수	계급값 x 상대도수
30	5		
50	10		
70	15		
90	40		
110	20		
130	10		
	합계 : 100		합계(평균값) :

정리

- 도수분포표에서의 평균값 계산
  - 평균값 = (계급값 x 상대도수)의 합계
- 히스토그램에서 평균값의 의미
  - 히스토그램을 지렛대라고 가정했을 때 평균값은 균형을 이루는 지점이다.
- 평균값의 성질
  - 데이터는 평균값 주변에 분포한다.
  - 많이 나타나는 데이터가 평균값에 주는 영향력은 크다.
  - 히스토그램이 좌우 대칭인 경우, 그 대칭축을 지나는 점이 평균값이 된다.

분산과 표준편차

- **불규칙한 통계량을 아는 것이 중요**
  - 평균값의 한계 → 데이터의 분포 중에서 하나의 수를 꺼낸 것에 불과하며, 데이터가 그 주변에 어느 정도 퍼져 있는지, 또는 흩어져 있는지는 알 수 없다.
  - 예)한 나라의 소득의 분포 → 균등한 소득분배 or 큰 빈부격차
  - 예)현재 버스의 운행 상황
  - 때로는 평균값보다 불규칙한 상태의 통계량을 아는 것이 중요할 때도 있다.

분산과 표준편차 (Cont.)

- **버스 도착시간으로 분산을 이해**
  - 7시 30분에 도착하는 버스가 5일 동안 도착한 시간

32	27	29	34	33

  - 평균값 : 31분 → 평균 이 버스는 7시 31분에 도착한다.
  - 버스가 도착한 시간은 제 각각이다.
  - 어느 정도 제 각각일까?
  - 이것을 어떻게 측정할까?

분산과 표준편차 (Cont.)

- **버스 도착시간으로 분산을 이해**
  - 5개의 각 데이터에서 평균값을 빼는 방법

+1	-4	-2	+3	+2

  - 각 데이터가 평균값으로부터 어느 정도 큰가, 또는 작은가를 나타낸다.
  - 통계학에서 이 각각의 수치를 **편차(Deviation)**이라 한다.
  - 도착 시간의 편차

+1	-4	-2	+3	+2

분산과 표준편차 (Cont.)

- **버스 도착시간으로 분산을 이해**
  - 5개의 편차를 축약하고, 하나의 수로 대표시키기
    - {(+1)+(-4)+(-2)+(+3)+(+2)} ÷ 5 =0
  - 어떤 데이터든지 그 편차를 만들어서 그 편차들을 산술평균으로 구하면 0이 된다. → 우리가 원하는 대표값이 아니다.
  - +와 –가 상쇄되지 않게 평균을 계산하는 방법 필요 → 제곱평균
  - 제곱평균 : 평균을 구하고 싶은 수치들을 각각 곱하고 모두 합하여 총 개수로 나눈 뒤에 루트를 하는 방법
  - 이렇게 함으로 최대값과 최소값 사이에 있는 어떤 하나의 수치를 산출가능 → 서로 상쇄되는 일이 없어진다.

분산과 표준편차 (Cont.)

- **버스 도착시간으로 분산을 이해**
  - 분산(Variance) 구하기
    - $\{(+1)^2 + (-4)^2 + (-2)^2 + (+3)^2 + (+2)^2\} \div 5$

      $= \{(+1)(+1) + (-4)(-4) + (-2)(-2) + (+3)(+3) + (+2)(+2)\} \div 5$

      $= \{1 + 16 + 4 + 9 + 4\} \div 5$

      $= 6.8$
  - 분산은 데이터가 퍼져 있는 상태를 평가할 수 있는 통계량
    - 하지만, 흩어져 있는 상태를 나타내는 수치로는 너무 크다(> ± 4).
    - 단위가 바뀐다(분minutes → 분^2).
    - 해결점은?

분산과 표준편차 (Cont.)

- **버스 도착시간으로 분산을 이해**
  - 분산에 루트를 씌어서 제곱평균을 구하면 해결
  - 표준편차(Standard Deviation) 구하기
    - $\sqrt{6.8} \fallingdotseq 2.61$

분산과 표준편차 (Cont.)

- **표준편차의 의미**
  - 버스는 평균적으로 시간표(7시 30분)보다 1분 늦는 버스다.
  - 그러나 이것을 아는 것만으로는 버스가 언제 올 지 알 수 없다. 버스는 언제나 1분 늦게 도착하는 것이 아니라 도착 시간이 제 각각이다.
  - 버스가 도착하는 시간의 불규칙성, 시간표와 맞지 않아서 확실하지 않은 상태를 측정하는 것이 표준편차이다. 그래서 계산한 값이 약 2.6분이다.
  - 이것은 무엇을 의미하는가?

분산과 표준편차 (Cont.)

- **표준편차의 의미**
  - **버스는 평균적으로 시간표보다 1분 늦게 도착하지만, 실제 도착 시간은 정해진 시간보다 전후로 대략 2.6분 정도 다를 수 있다.**
  - 평균값은 데이터의 분포를 대표하는 수치이고, 표준편차는 그 대표값을 기점으로 해서 데이터가 대략 어느 정도 멀리까지 위치해 있는지를 나타내는 통계량이다.

분산과 표준편차 (Cont.)

• 표준편차의 의미

- 10점 만점인 시험에서 받은 결과

X데이터	4	4	5	6	6
Y데이터	1	2	6	7	9

평균값 = 5

평균값 = 5

- 두 점수의 데이터 편차

X데이터	-1	-1	0	+1	+1
Y데이터	-4	-3	+1	+2	+4

분산과 표준편차 (Cont.)

- **표준편차의 의미**
  - X데이터의 표준편차
    - $\sqrt{\{(-1)^2 + (-1)^2 + (0)^2 + (+1)^2 + (+1)^2\} \div 5} \fallingdotseq 0.89$
  - Y데이터의 표준편차
    - $\sqrt{\{(-4)^2 + (-3)^2 + (+1)^2 + (+2)^2 + (+4)^2\} \div 5} \fallingdotseq 3.03$
  - 결국 Y데이터의 표준편차가 크다.

분산과 표준편차 (Cont.)

- 도수분포표로 표준편차를 구하는 방법
  - {(계급값 - 평균값)^2 x (상대도수)}의 합계 = 분산
  - $\sqrt{\text{분산}}$ = 편차의 제곱평균

A(계급값)	B(상대도수)	A X B	C(계급값 – 평균값)	C^2	B(상대도수)	C^2 x B
1	0.3	0.3	-1	1	0.3	0.3
2	0.5	1.0	0	0	0.5	0
3	0.1	0.3	+1	1	0.1	0.1
4	0.1	0.4	+2	4	0.1	0.4

평균값 = 2.0

분산 = 0.8

표준편차 = $\sqrt{0.8}$ ≒ 0.89

연습문제

- **다음에 나타낸 데이터의 표준편차를 다음의 순서대로 계산하시오.**

1. 우선 평균을 구하시오.

데이터	6	4	6	6	6	3	7	2	2	8	평균값 :

2. 편차를 계산하시오.

편차										

연습문제

- **다음에 나타낸 데이터의 표준편차를 다음의 순서대로 계산하시오.**

3. 편차의 제곱과 그 평균(=분산)을 계산하시오.

편차의 제곱											평균값 :

4. 표준편차를 계산하시오.

표준편차 = (편차의 제곱평균)의 제곱근($\sqrt{\quad}$)=

정리

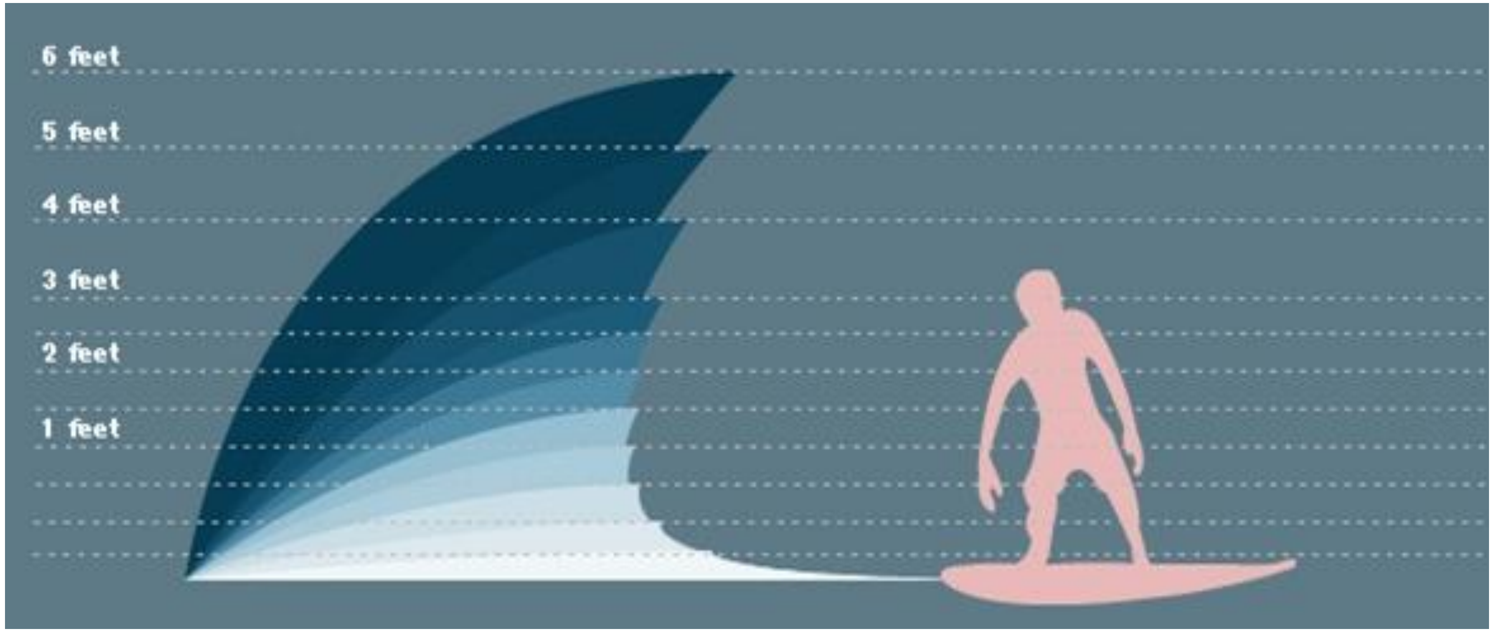
- **평균값 계산**
  - (데이터의 총합) ÷ (데이터의 총 개수)
- **편차 계산**
  - (데이터의 수치) – (평균값)
- **분산 계산**
  - {(편차 제곱)의 총합} ÷ (데이터의 총 개수)
- **표준편차 계산**
  - $\sqrt{\text{분산}}$ = 편차의 제곱평균
- **도수분포표를 이용해 계산하는 분산과 표준편차**
  - 분산 = {(계급값 – 평균값)^2 x (상대도수)}의 합계
  - 표준편차 = $\sqrt{\text{분산}}$

정리 (Cont.)

- **표준편차의 의미**
  - 평균값은 분포하고 있는 데이터 중에서 대표적인 수로 꺼낸 것
  - 데이터는 평균값을 기점으로 그 앞뒤에 널리 펴져 있다.
  - 그러나 어느 정도 퍼져 있거나 흩어져 있는지는 평균값으로 알 수 없다.
  - 퍼져 있거나 흩어져 있는 정도를 평가하는 것이 표준편차다.
  - 표준편차는 데이터들의 평균값에서 떨어져 있는 것을 평균화한 것
  - 이때 멀리 떨어져 있든지 가까운 곳에 있든지, 모두 양수로 평가하여 상쇄되지 않도록 해서 평균을 구한다.

표준편차(1)

- **표준편차는 '파도의 거칠기'**
  - 바다의 수위 → 평균값
  - 파도가 거칠게 쳐서 수위의 차가 커지는 것 → 표준편차



표준편차(1) (Cont.)

- **표준편차로 데이터의 특수성을 평가**
  - 표준편차를 알면 무엇을 알 수 있는가?
    - 한 데이터 세트 중에 있는 어떤 데이터 하나의 수가 갖는 의미
    - 여러 데이터 세트들을 서로 비교해서 나타나는 차이
  - 시험점수가 평균 75점으로, 평균 점수인 60점보다 15점이 높다면, 나는 과연 얼만큼의 기쁨을 갖게 될 것인가?
  - 즉, 표준편차가 몇 점인가?

표준편차(1) (Cont.)

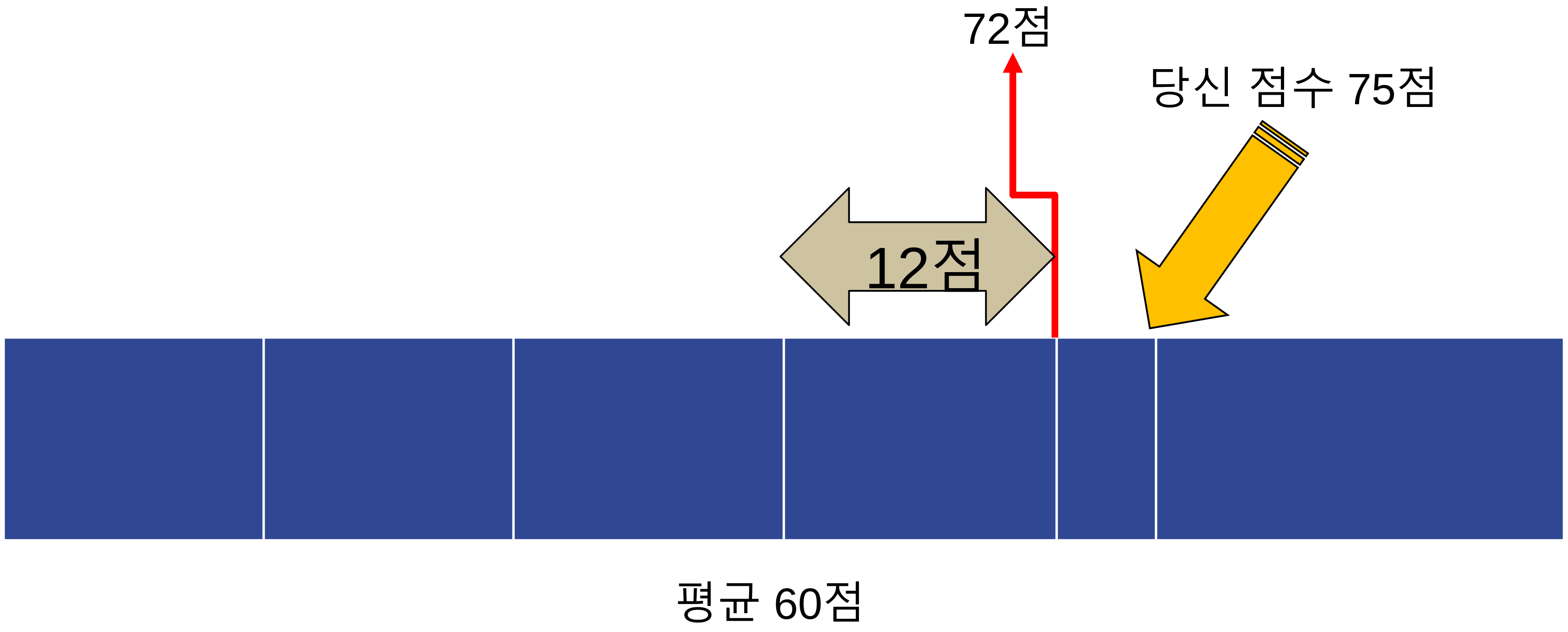
- **표준편차로 데이터의 특수성을 평가**
  - 만일 표준편차가 12점이라면?
  - 내가 받은 점수는 대략 표준편차만큼 더 높은 점수라는 의미.
  - 그렇다면 나의 점수는 평균점수보다 잘한 쪽(평균보다 높은 쪽)에서 보통으로 떨어져 있는 점수이다.
  - 즉, 일반적으로 떨어져 있는 정도의 점수.
  - 다시 말해서, 이 정도의 점수를 받은 사람이 많다는 뜻
  - 다른 말로 말하면...그리 될 듯이 기쁘지는 않다는 뜻...

표준편차(1) (Cont.)

- **표준편차로 데이터의 특수성을 평가**
  - 만일 표준편차가 8점이라면?
  - 이것은, 내가 받은 점수가 평균점수에서 표준편차의 2개 정도나 떨어져 있다는 뜻.
  - 훨씬 기분이 좋아야 한다는 뜻.

표준편차(1) (Cont.)

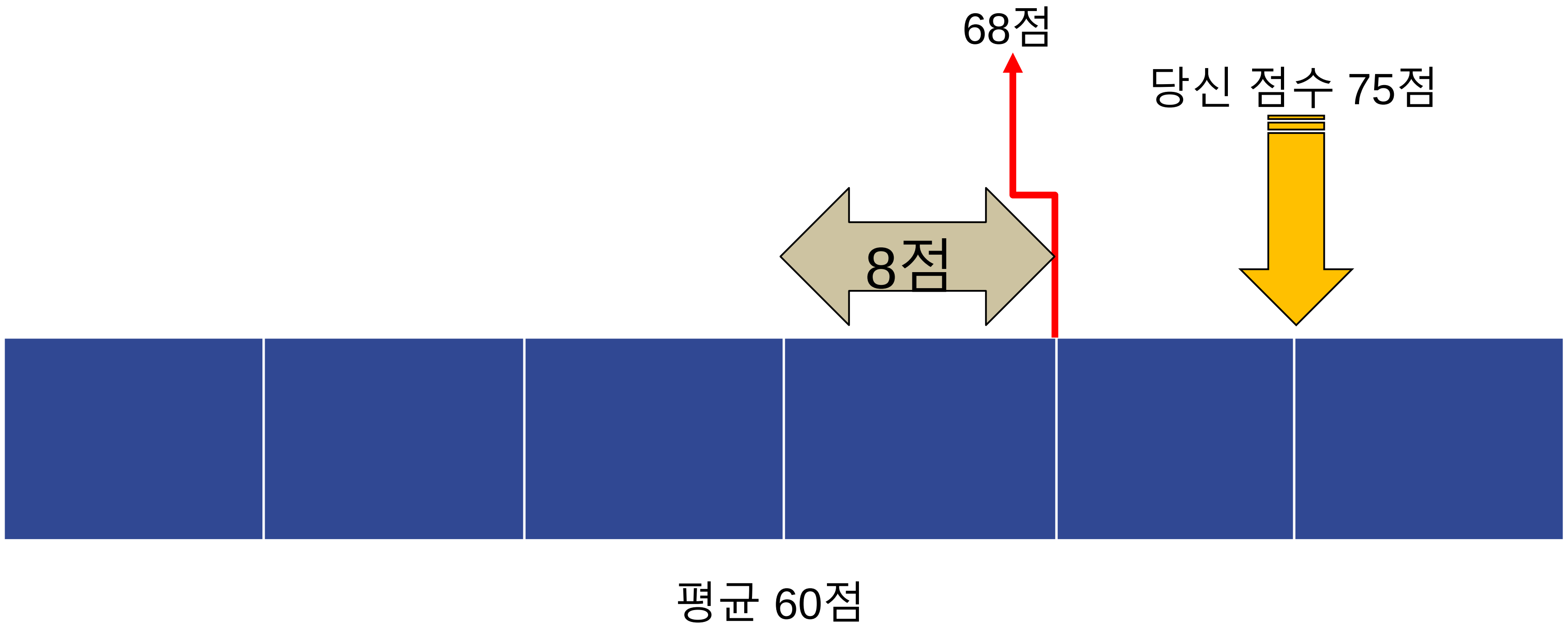
- **표준편차로 데이터의 특수성을 평가**
  - 표준편차가 12점인 경우



  - 표준편차만큼 떨어져 있지 않다. → **보통 성적**

표준편차(1) (Cont.)

- **표준편차로 데이터의 특수성을 평가**
  - 표준편차가 8점인 경우



  - 표준편차의 약 2배나 멀리 떨어져 있다. → **좋은 성적**

표준편차(1) (Cont.)

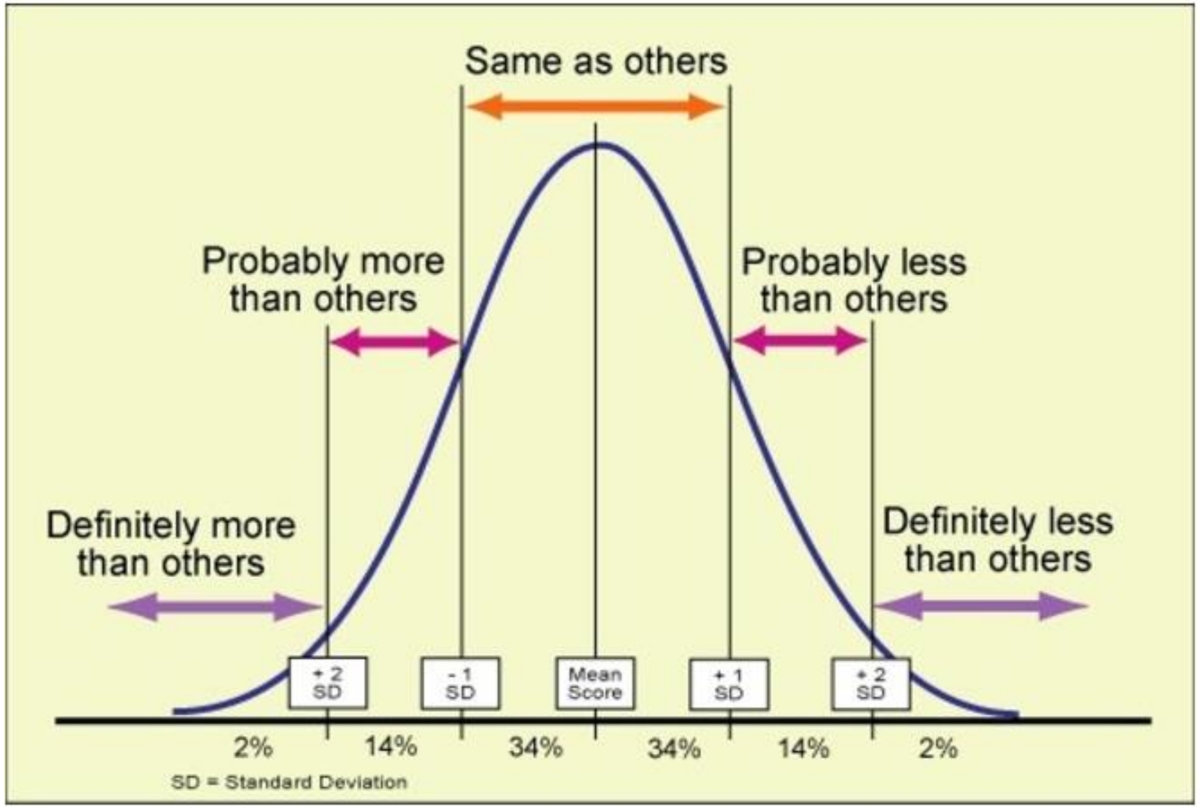
- **표준편차로 데이터의 특수성을 평가**
  - 결론 : 한 데이터 세트 중에 있는 어떤 하나의 데이터가 가진 특수성은 평균에서 떨어진 정도(=편차)를 나타내는 수치만으로는 계측할 수 없고, 표준편차를 기준으로 가정해야만 알 수 있다.
  - 그래서, 편차를 표준편차로 계산해서 **얼마만큼**이라고 나타내는 변환이 중요하다.

표준편차(1) (Cont.)

- **데이터의 특수성의 평가 기준**
  - 데이터 세트 중에 있는 어느 한 데이터의 편차가 표준편차로 계산해서 ±1배 전후라면 이것은 ***평범한 데이터***라고 할 수 있으며, ±2배로 멀리 있는 멀리 있는 데이터일 경우는 ***특수한 데이터***라고 할 수 있다.
  - 여기서 특수하다는 말이 어느 정도를 뜻하는가?
  - 만일 데이터의 성질이 좋다면, 즉 정규분포에 가깝다고 한다면…평균값에서 표준편차 ±1배의 범위 내에 약 70%의 데이터가 들어간다.
  - 표준편차 ±2배보다 멀리 떨어진 데이터는 좌우 양쪽을 합쳐서 5%밖에 없다는 의미.

표준편차(1) (Cont.)

- 데이터의 특수성의 평가 기준
  - 그렇다면 앞의 경우처럼 만일 당신의 데이터(점수)가 평균값보다 큰 쪽으로 표준편차의 2배 이상 떨어져 있다면, 그것은 무슨 뜻?
  - **전체의 2.5% 범위 내에 있는 데이터**
  - 즉...상당히 그 데이터는 **특수한 경우**에 있다는 뜻.



표준편차(1) (Cont.)

- **여러 데이터 세트를 비교할 때의 표준편차**
  - X군은 10번 모의시험을 본 평균점수가 60점이고, 표준편차가 10점이라고 하자.
  - Y군은 X군과 같은 모의시험을 10번 본 평균점수가 50점이고, 표준편차가 30점이라고 하자.
  - 이것으로 무엇을 읽어 낼 수 있을까?

표준편차(1) (Cont.)

- **여러 데이터 세트를 비교할 때의 표준편차**
  - 평균점수만 보면 X군이 Y군보다 공부를 잘하는 학생이지만, 이것만으로는 이 두 사람이 진짜 시험을 치렀을 때 얻을 점수를 예측할 수 없다.
  - X군의 평균 점수는 60점, 표준편차는 10점이기 때문에 X군은 표준편차 ±1배 정도의 폭, 대략 50 ~ 70점 범위의 점수를 맞는 학생이라고 판단할 수 있다.
  - Y군은 평균 점수가 50점, 표준편차가 30점이기 때문에 대략 20 ~ 80 점 범위의 점수를 맞는 학생이라고 판단할 수 있다.
  - 즉, X군은 안정된 점수를 맞는 학생이고, Y군은 시험을 볼 때마다 점수 차가 큰 학생이라고 말할 수 있다.

표준편차(1) (Cont.)

- **여러 데이터 세트를 비교할 때의 표준편차**
  - 그렇다면 X군이 Y군보다 공부를 잘 하기 때문에 더 좋은 학교에 갈 수 있는가?
  - 결코 그렇게 단언할 수 없다.
  - X군은 50점을 맞으면 갈 수 있는 학교에는 합격할 수 있지만, 80점 커트라인 학교에는 상당히 들어가기 힘들다.
  - Y군은 40점으로 갈 수 있는 학교에도 떨어질 수 있지만, 80점 커트라인 학교에 합격할 수 있다.
  - 즉, X군과 Y군은 공부를 잘하는 것이라는 서열적인 평가가 아니라 성질이 다른 것으로 평가할 수 있다는 의미이다.

표준편차(1) (Cont.)

- **가공된 데이터의 평균값과 표준편차**
  - X데이터 1, 3, 4, 5, 7
  - 각 수에 4를 더한다
  - Y데이터 5, 7, 8, 9, 11
  - X데이터의 평균값 {1 + 3 + 4 + 5 + 7} ÷ 5 = 4
  - Y데이터의 평균값 {5 + 7 + 8 + 9 + 11} ÷ 5 = 8
  - 즉 Y데이터의 평균값은 X의 평균에 미리 더한 4만큼 커진다.

표준편차(1) (Cont.)

- **가공된 데이터의 평균값과 표준편차**
  - X데이터의 편차 -3, -1, 0, +1, +3
  - Y데이터의 편차 -3, -1, 0, +1, +3 → 같다
  - X데이터의 분산 $\{(-3)^2+(1)^2+0^2+(+1)^2+(+3)^2\} \div 5 = 4$
  - Y데이터의 분산 $\{(-3)^2+(1)^2+0^2+(+1)^2+(+3)^2\} \div 5 = 4$
  - X데이터의 표준편차

    $\sqrt{4} = 2$
  - Y데이터의 표준편차

    $\sqrt{4} = 2$ → 같다.

표준편차(1) (Cont.)

- **가공된 데이터의 평균값과 표준편차**
  - X데이터에 있는 각 수에 4를 더하는 가공을 하여 Y데이터를 만들었다.
  - Y데이터는 X데이터에 비해 평균값이 4만큼 커진다.
  - 이것은 모든 데이터가 4만큼 증가했기 때문이다.
  - 히스토그램이 오른쪽으로 4만큼 이동했기 때문에 지렛대가 균형을 이루는 지점도 이와 같이 이동한다는 의미이다.
  - 그러면 서로 편차도 같아지게 된다.

표준편차(1) (Cont.)

- 가공된 데이터의 평균값과 표준편차
  - X데이터의 모든 수에 일정한 수 a를 더해서 새로운 Y데이터를 만들면, Y데이터의 평균값은 X데이터의 평균값에 a를 더한 것이 되며, Y데이터의 분산과 표준편차는 원래의 X데이터 수치와 같다.

표준편차(1) (Cont.)

- **가공된 데이터의 평균값과 표준편차**
  - X데이터 1, 3, 4, 5, 7
  - 각각에 2배를 한다.
  - Y데이터 2, 6, 8, 10, 14
  - X데이터의 평균값 {1 + 3 + 4 + 5 + 7} ÷ 5 = 4
  - Y데이터의 평균값 {2 + 6 + 8 + 10 + 14} ÷ 5 = 8
  - 즉 Y데이터의 평균값은 X의 평균에 2배가 된다.

표준편차(1) (Cont.)

- **가공된 데이터의 평균값과 표준편차**
  - X데이터의 편차 -3, -1, 0, +1, +3
  - Y데이터의 편차 -6, -2, 0, 2, 6 → X데이터의 편차의 2배가 된다.
  - X데이터의 분산 $\{(-3)^2+(1)^2+0^2+(+1)^2+(+3)^2\} \div 5 = 4$
  - Y데이터의 분산 $\{(-6)^2+(-2)^2+0^2+(+2)^2+(+6)^2\} \div 5 = 16$ → 4배가 된다.
  - X데이터의 표준편차

    $\sqrt{4} = 2$
  - Y데이터의 표준편차

    $\sqrt{16} = 4$ → 2배가 된다.

표준편차(1) (Cont.)

- **가공된 데이터의 평균값과 표준편차**
  - X데이터에 있는 각 수에 2를 곱하는 가공을 하여 Y데이터를 만들었다.
  - Y데이터는 X데이터에 비해 평균값이 2배가 된다.
  - 편차도 2배가 된다.
  - 이것으로 분산은 2의 제곱배로 4배가 된다는 것을 알 수 있다.
  - 그래서 표준편차는 2배가 된다.

표준편차(1) (Cont.)

- 가공된 데이터의 평균값과 표준편차
  - X데이터의 모든 수에 일정한 수 k를 곱해서 새로운 Y데이터를 만들면, Y데이터의 평균값은 X데이터의 평균값에 k를 곱한 것이 되며, Y데이터의 분산은 k의 제곱배, 표준편차는 k배가 된다.

연습문제

- **괄호 안을 채우고, 올바른 것에 o표 하시오.**
  - 성인 여성의 키 평균값을 160cm, 표준편차를 10cm라고 할 때

① 키가 150cm의 여성은 표준편차로 계산해서 (   )배 정도 평균값보다 낮다. 이것은 데이터로 봤을 때 특수하다고(말할 수 있다, 말할 수 없다).

② 키가 185cm의 여성은 표준편차로 계산해서 (   )배 정도 평균값보다 높다. 이것은 데이터로 봤을 때 특수하다고 (말할 수 있다, 말할 수 없다).

정리

- 데이터의 특수성을 판단하는 데는 표준편차를 기준으로 한다.
- 평균에서 표준편차 1배 정도 떨어져 있는 데이터는 평범한 데이터라고 할 수 있다. 또한 평균에서 표준편차 2배 이상 떨어져 있는 데이터는 특수한 데이터라고 할 수 있다.
- 표준편차의 얼마만큼 이라는 것을 알기 위해서는 {(데이터) – (평균값)} ÷ (표준편차) 를 계산하면 된다.

정리

- X데이터의 모든 수에 일정한 수 a를 더해서 새로운 Y데이터를 만들면, Y데이터의 평균값은 X데이터의 평균값에 a를 더한 것이 되며, Y데이터의 분산과 표준편차는 원래의 X데이터 수치와 같다.
- X데이터의 모든 수에 일정한 수 k를 곱해서 새로운 Y데이터를 만들면, Y데이터의 평균값은 X데이터의 평균값에 k를 곱한 것이 되며, Y데이터의 분산은 k의 제곱배, 표준편차는 k배가 된다.
- 데이터를 {(데이터) – (평균값)} ÷ (표준편차)로 가공하면, 이 데이터로 구한 평균값은 0이고, 표준편차는 1이 된다.

표준편차(2)

- **주식거래에서 이익을 남기기 위해서 어떻게 해야 할까?**
  - 배당을 받고 이것을 수익으로 하는 것 → Income Gain
  - 주식을 싸게 사서 비쌀 때 팔아 그 차액을 수익으로 남기는 것 → Capital Gain
- **Capital Gain을 목적으로 주식을 거래할 경우 중요해지는 것은 주식의 평균수익률이다.**
- **월평균수익률 : 어느 회사의 주식이 1개월 동안에 몇 % 상승 또는 하락(마이너스 상승)했는가를 연 12개월에 걸친 데이터로 수집해 평균을 구한 것**

표준편차(2) (Cont.)

- 월평균수익률
  - 예) 월평균수익률이 10%이다.
  - 이 회사의 주식이 평균적으로 1개월에 10% 상승했다는 것을 의미.
  - 이 주식을 100만원어치 구입해서 1개월 동안 보유한 뒤 매각하면, 평균 10% 상승한 10만원을 수익으로 남길 수 있다는 의미.

표준편차(2) (Cont.)

- **평균수익률만으로는 우량기업인지 판단할 수 없다.**
  - 다음은 한 기업의 주식 월평균수익률이다.

연도	1980	1981	1982	1983	1984	평균
월평균수익률	2.05	2.46	-1.33	2.04	-0.54	**0.94**

  - 1981년도 월평균수익률은 대략 2.5%이다.
  - 이것만 보면, 그 해의 주식거래에서 상당한 이익을 남겼을 것이다.
  - 한 달 사이에 수익률이 2.5%라는 것은 한 해 12를 곱해서 30%의 이익이 발생한다는 의미.
  - 즉, 예금을 100만원 저축하면 1년 뒤에 30만원 이자(단, 단리에 해당)가 붙어서 130만원이 된다는 뜻.

표준편차(2) (Cont.)

- **평균수익률만으로는 우량기업인지 판단할 수 없다.**
  - 하지만,
  - 이 사실만으로 투자한다면 절대로 안 된다.
  - 왜냐하면 이것은 어디까지나 **평균값**이라는 것이다.
  - 수익의 평균값이 2.5%라고 해도 매월 2.5%씩 수익을 올릴 수 있는 것은 아니다.
  - 실제로 올릴 수 있는 수익은 그 값을 기점으로 해서 그 앞뒤에 해당하는 값이다.

표준편차(2) (Cont.)

	1980	1981	1982	1983	1984
1월	9.2	2.8	-0.6	-2.8	0
2월	2.3	-1.4	-11.8	9.3	-5.7
3월	-6.5	17.6	3.5	11.4	10.6
4월	9	17.8	1.9	3	-0.6
5월	5.3	5.5	-5.5	-7.5	-11.2
6월	-4.3	-1.9	-9.1	2.5	-3.8
7월	-3.7	1.9	-5.7	-0.6	-5.2
8월	7	9	2.3	1.8	6.2
9월	7.6	-10.3	-4.9	5.1	-4.2
10월	1.4	-10.3	-0.8	-2.3	2.1
11월	-3.4	-7.7	8	-6	0.6
12월	0.7	6.5	6.7	10.6	4.7

주식의 월별수익률

표준편차(2) (Cont.)

- **평균수익률만으로는 우량기업인지 판단할 수 없다.**
  - 1981년의 데이터를 보면 실제로 월별수익률은 다양하다.
  - 오히려 평균값 2.5%에 근접하는 숫자는 1월 밖에 없다.
  - 이럴 때,
  - 데이터의 실제 상황을 조금 더 자세히 파악할 수 있는 통계량이 필요하다.
  - 그것이 바로 **표준편차**이다.
  - 앞의 표에서 보듯 1981년 월평균수익률이 약 2.5%이지만, 표준편차는 9%를 넘는다.

표준편차(2) (Cont.)

- **평균수익률만으로는 우량기업인지 판단할 수 없다.**
  - 따라서,
  - 1981년 월별수익률은 2.5 ± 9.0%의 범위, 즉 +11.5 ~ -6.5 범위의 수익률은 보통으로 관측된다.
  - 다시 말하면,
  - 월평균 2.5%의 수익을 올리는 주식을 살 때, 6.5%의 손실을 볼 수 있다는 점을 각오해야 한다는 의미.

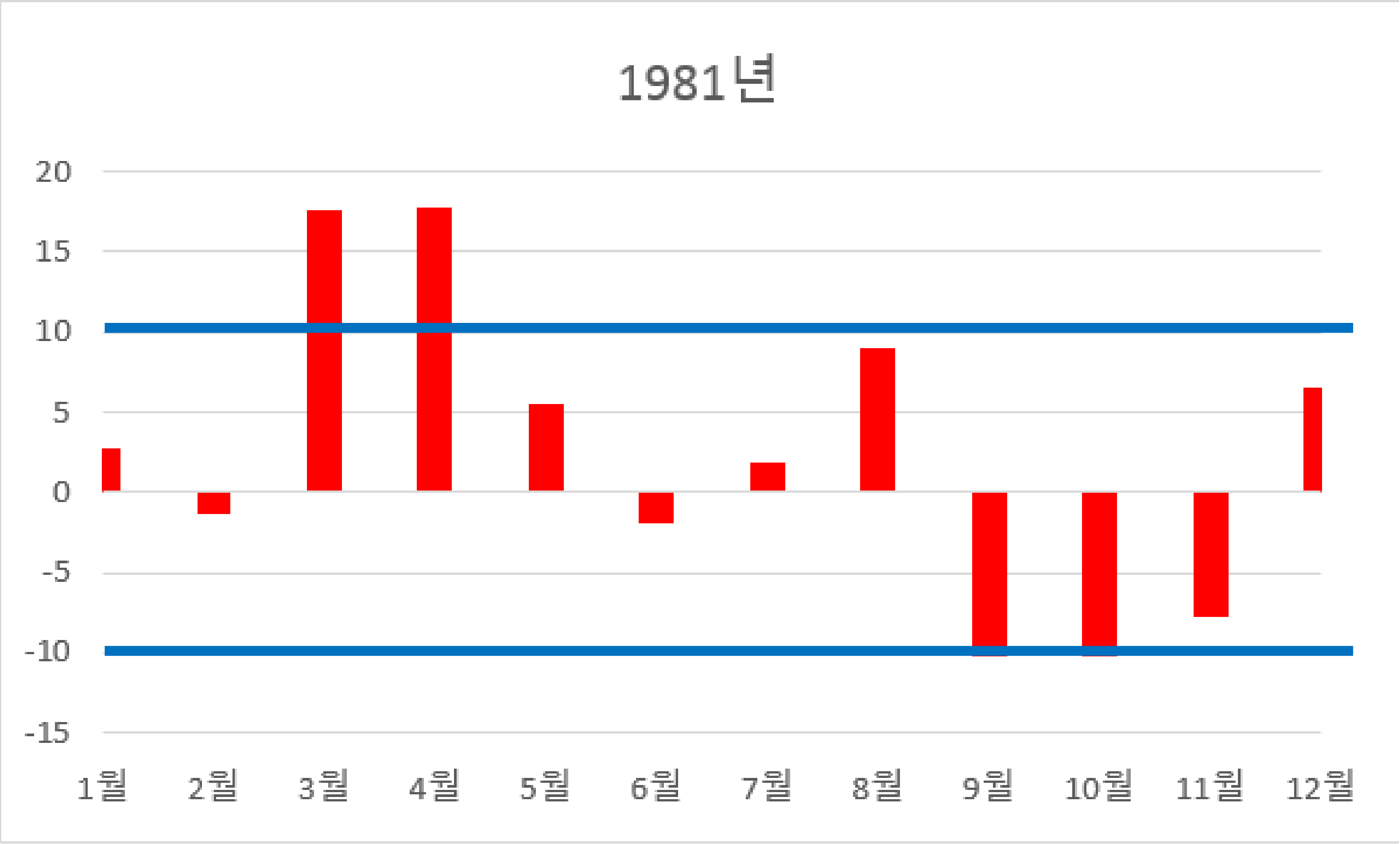
연도	1980	1981	1982	1983	1984	평균
월평균수익률	2.05	2.46	-1.33	2.04	-0.54	0.94
표준편차	5.35	9.11	5.91	5.98	5.71	6.74

표준편차(2) (Cont.)

- 주가변동성이 의미하는 것
  - 1981년의 월별수익률을 막대그래프로 나타내면,
  - 평균값을 기점으로 위와 아래로 물결치고 있다(파도가 치는 바다에서의 서핑을 생각해보라)는 것을 알 수 있다.
  - 각각 다른 파도의 높이를 일반적으로 본 폭이 표준편차이다.
  - 그래프에서 평균값으로부터 아래로 표준편차만큼 내려간 곳과 위로 올라간 곳에 선을 그려보면 대부분의 막대기가 그 범위 안에 있다는 것을 볼 수 있다.

표준편차(2) (Cont.)

- 주가변동성이 의미하는 것



표준편차(2) (Cont.)

- **주가변동성이 의미하는 것**
  - 이렇게 주식거래에서는,
  - 수익률의 평균값만이 아니라 그 표준편차도 중요하다.
  - 그렇기 때문에 주식에서는 이 표준편차를 뜻하는 전문용어가 있는데, 그것을 **주가변동성(Volatility)**이라고 한다.
  - 즉,
  - **평균값에서 어느 정도의 폭으로 변동이 생기는가를 의미하는 말**

표준편차(2) (Cont.)

- **주가변동성이 의미하는 것**
  - 그래서,
  - 주식 수익률의 표준편차 = 주가변동성은 주식거래 리스크의 지표라고 생각할 수 있다.
  - 수익으로 그 평균값을 예상해도 그 값에서부터 주가변동성만큼 떨어지는 경우도 충분히 예상해야 하기 때문.
  - 결론은,
  - **주가변동성은 바로 위험성을 나타내는 지표**이다.
  - 그러나,
  - 9% 정도 하락할 수 있다는 것은 곧 9% 수익도 된다는 의미이기 때문에 리스크를 나타내는 지표이기도 하지만, 이것은 기회를 나타내는 지표가 되기도 한다.

표준편차(2) (Cont.)

- **주가변동성이 의미하는 것**
  - 결론은,
  - **주가변동성이 9%라면 평균값에서 (표준편차 x 2 =)18% 이상 떨어지는 일은(물론 올라가는 일도) 거의 없을 것이라고 생각해도 된다는 의미**

연습문제

- **1983년 한 기업의 주식에 투자했을 때, 월평균수익률은 약 2%, 표준편차는 6%였다.**
  ① 이 해 투자는 월평균으로 투자액의 2%를 기대할 수 있지만, 전후로 표준편차 1배 정도의 변동은 평균적으로 일어난다고 생각해야 한다. 다시 말해, 2% - (   )% ~ 2% + (   )%으로 계산하고, (   )% - (   )%의 변동 폭으로 달라질 것이라는 생각을 미리 해둘 필요가 있다.
  ② 일반적으로는 표준편차의 2배 정도 오르거나 떨어질 경우는 별로 생각하지 않아도 된다. 즉, 월간 수익률이 2% + (   ) x 2 = (   )%가 되거나, 2% - (   ) x 2 = (   )%가 되는 경우는 드물다고 생각해도 좋다.

연습문제

- 주식 A는 월평균수익률이 7%이고, 표준편차는 12%다. 주식 B는 월평균수익률이 4%이고, 표준편차는 3%이다. 이때 주식 A를 사서 1개월 가지고 있을 때의 수익률은 (　　　)% ~ (　　　)%라고 예상할 수 있으며, 주식 B를 사서 1개월간 가지고 있을 때의 수익률은 (　　　)% ~ (　　　)%라고 예상할 수 있다.
- 그래서 원금손실을 바라지 않는 투자자는 주식 (　　　)를 구입해야 하며, 이 경우 좋은 성적을 거둘 때의 수익은 반드시 (　　　)% 정도라고 생각해야 한다. 반대로 원금손실을 두려워하지 않는 투자자는 주식 (　　　)를 구입해야 하며, 이 경우 운이 좋은 경우 (　　　)% 정도의 수익은 충분히 얻을 수 있을 것이라고 예상된다.

정리

- 주식거래의 지표는 수익률의 평균값뿐만 아니라 표준편차도 중요하다.
- 주식에 투자할 때는 수익률의 평균값이 표준편차 1배 정도 떨어진 수익률이 될 경우도 각오해 두는 것이 좋다.
- 주식에 투자할 때는 수익률의 평균값이 표준편차 2배 정도 떨어진 수익률이 될 경우는 거의 없을 것이라고 생각해도 된다.
- 주식 수익률의 표준편차를 전문용어로 주가변동성이라고 한다.

표준편차(3)

- **High Risk, High Return, Low Risk, Low Return**
  - 앞 강의에서, 주식 수익률의 표준편차는 주가변동성이라고 하며, 주식거래의 **리스크**를 나타내는 것
  - 수익률의 표준편차가 큰 주식은 평균에서 표준편차 1배 정도 수익률이 낮아지는 것이 일반적인 현상이기 때문에 이것은 **위험성(리스크)**이라고 인식해야 한다.
  - 다양한 자산운용 방법들은 어느 정도의 수익률과 어느 정도의 주가변동성을 보이는가?

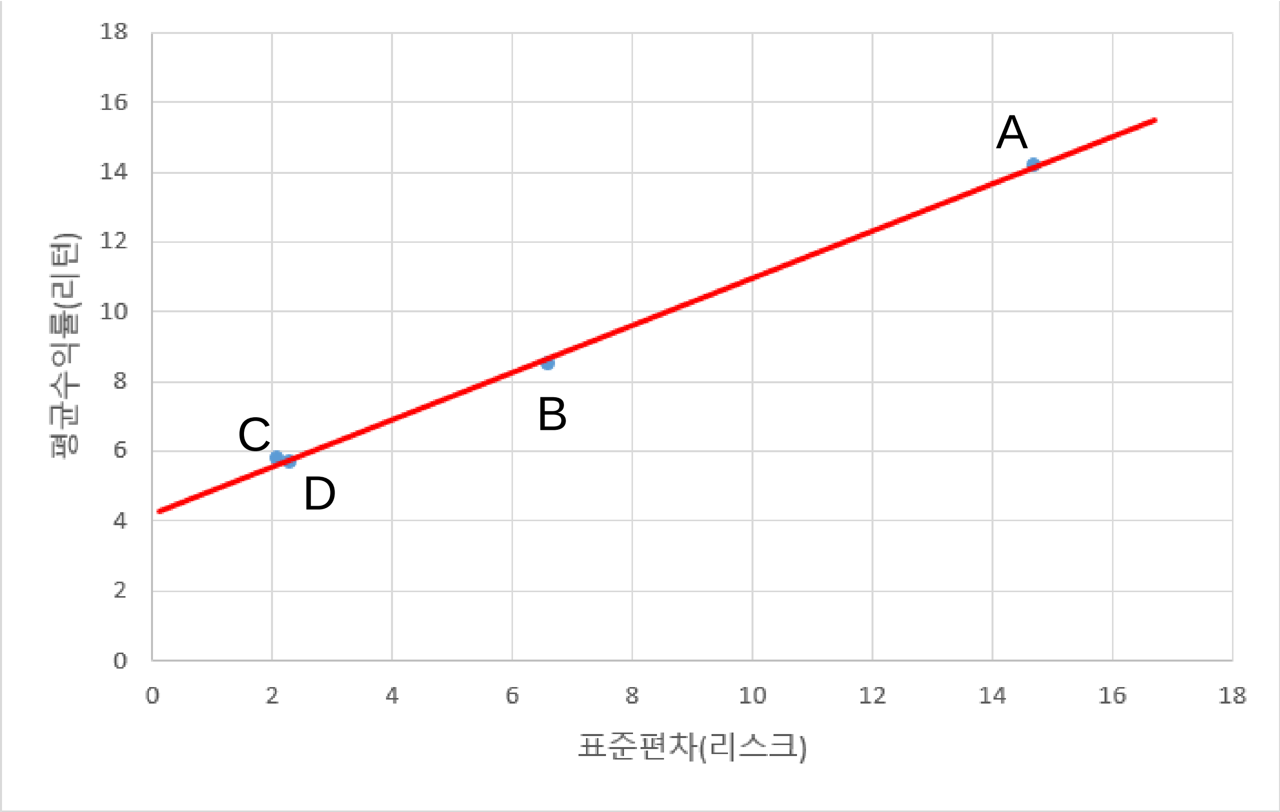
표준편차(3) (Cont.)

	주식펀드(상품A)	채권펀드(상품B)	MMMF(상품C)	1년 정기예금(상품D)
1988년	13.2	7.7	7.3	7.4
1989년	20.9	9.5	9	8.2
1990년	-6.9	3.7	8.1	7.9
1991년	35.6	17.2	5.9	7.1
1992년	8.9	7.9	3.3	4.2
1993년	12.5	10.3	2.6	3.3
1994년	-1.7	-3.7	3.8	3
1995년	31.1	15.6	5.4	4.9
리스크=표준편차	**14.7**	**6.6**	**2.3**	**2.1**
리턴=평균수익률	**14.2**	**8.5**	**5.7**	**5.8**

한 연구소에서 조사한 1988~1995년 사이의 미국 뮤추얼펀드 자산운용 실적

표준편차(3)(Cont.)

- **High Risk, High Return, Low Risk, Low Return**
  - 앞의 표를 보면, 평균수익률이 높은 운용은 표준편차도 크다.
  - 평균수익률(세로축의 값)이 높은 펀드는 표준편차(가로축의 값)도 크다.
  - 리스크(표준편차)를 작게 하려고 하면 평균수익률도 자동적으로 작아야만 한다.
  - High Risk, High Return



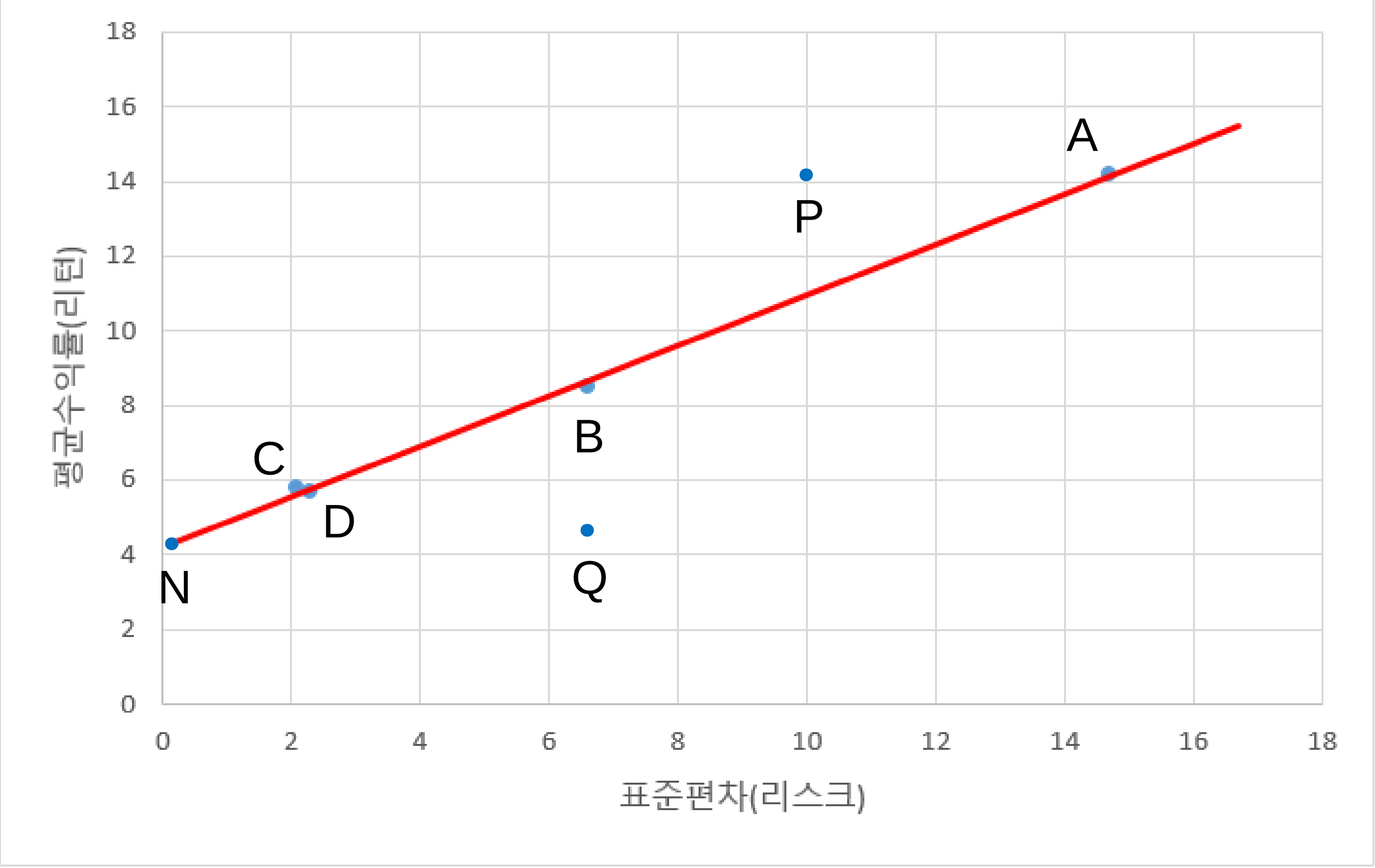
표준편차(3)(Cont.)

- **금융 상품의 우열을 가리는 방법**
  - High Risk, High Return, or Low Risk, Low Return은 각각 한 쌍이 되는 것으로, 어느 쌍이 어느 쌍에 비해 우수하다거나 열등하다고 말할 수 없다.
  - 그것은 투자자의 기호의 문제이다.
  - 어쩌면, 상품성은 같다고 볼 수 있다.

표준편차(3)(Cont.)

• 금융 상품의 우열을 가리는 방법

- 옆의 차트에서 보듯, A B C D의 금융상품 및 나머지 금융상품은 상품성에 우열이 없다고 생각해야 한다.



표준편차(3)(Cont.)

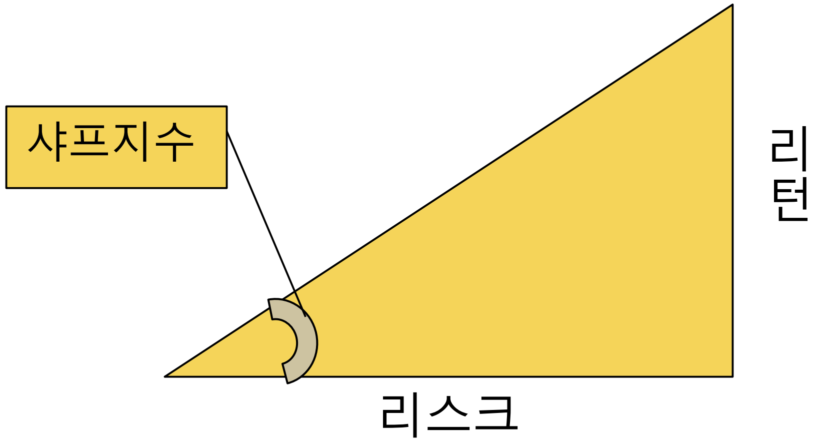
- **금융 상품의 우열을 가리는 방법**
  - P상품과 A상품을 비교해보라.
    - P상품의 리턴과 같은 A상품은 표준편차(리스크)에 차이가 난다.
    - 즉 리턴은 같지만, 리스크가 A상품보다 P상품이 낫다.
    - 따라서 P상품은 A상품보다 우수한 금융상품이다.
  - Q상품과 B상품을 비교해보라.
    - 서로 리스크는 같지만 Q상품의 리턴이 작다.
    - 따라서 Q는 B에 비해 열등한 금융상품이다.
  - 즉, 직선 상의 A,B,C,D보다 위에 있는 금융상품은 직선 위에 있는 어느 금융상품보다 뛰어난 상품이고, 반대로 그 아래에 있는 금융상품은 열등한 상품이다.

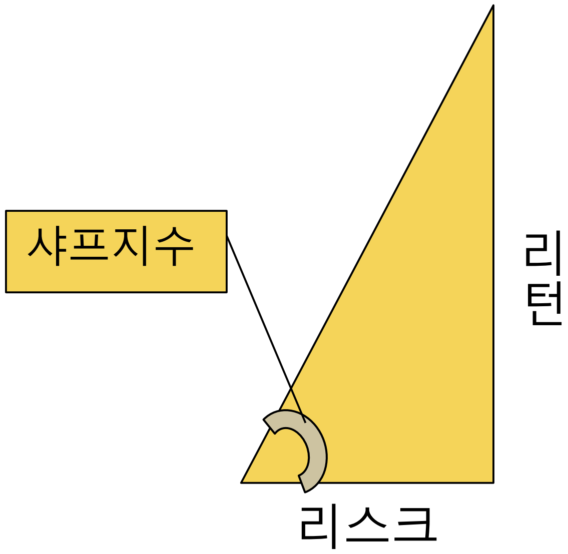
표준편차(3)(Cont.)

- **금융상품의 우열을 가리는 수치, 샤프지수**
  - 각 금융상품의 우열을 도표가 아닌 하나의 수치로 바꾸어 사용할 방법은 없을까?
  - 그 답은, 샤프라는 경제학자가 만든 샤프지수가 있다.
  - 샤프지수가 클수록 우량 금융상품으로 평가된다.

표준편차(3)(Cont.)

- **금융상품의 우열을 가리는 수치, 샤프지수**
  - (X의 샤프지수) = {(X의 리턴) – (국채 이자율)} ÷ (X의 리스크)
  - 샤프지수는 분수형태이다.
  - 분자는 리턴 평가, 분모는 리스크 평가를 나타낸다.
  - 따라서, 분자(리턴)가 크면 샤프지수도 커지고, 분모(리스크)가 작아져도 샤프지수는 커진다.





표준편차(3)(Cont.)

- **금융상품의 우열을 가리는 수치, 샤프지수**
  - 리턴에서 국채 이자율(무위험 수익률)을 빼는 이유는, 국채는 회사에 비해 파산할 가능성이 아주 적기 때문에 국채는 리스크가 적은 금융자산이기 때문이다.
  - 그 국채의 이자율을 웃도는 수익만큼을 리스크(표준편차)로 나누는 이유는, 똑같은 리턴을 얻더라도 리스크가 높은 금융상품은 운용 상태가 나쁜 상품이라고 판단되기 때문이다.
  - 즉, 리스크가 2인 경우에 수익은 절반으로 낮아지고, 리스크가 3인 경우에는 수익이 3분의 1로 낮아지게 된다.

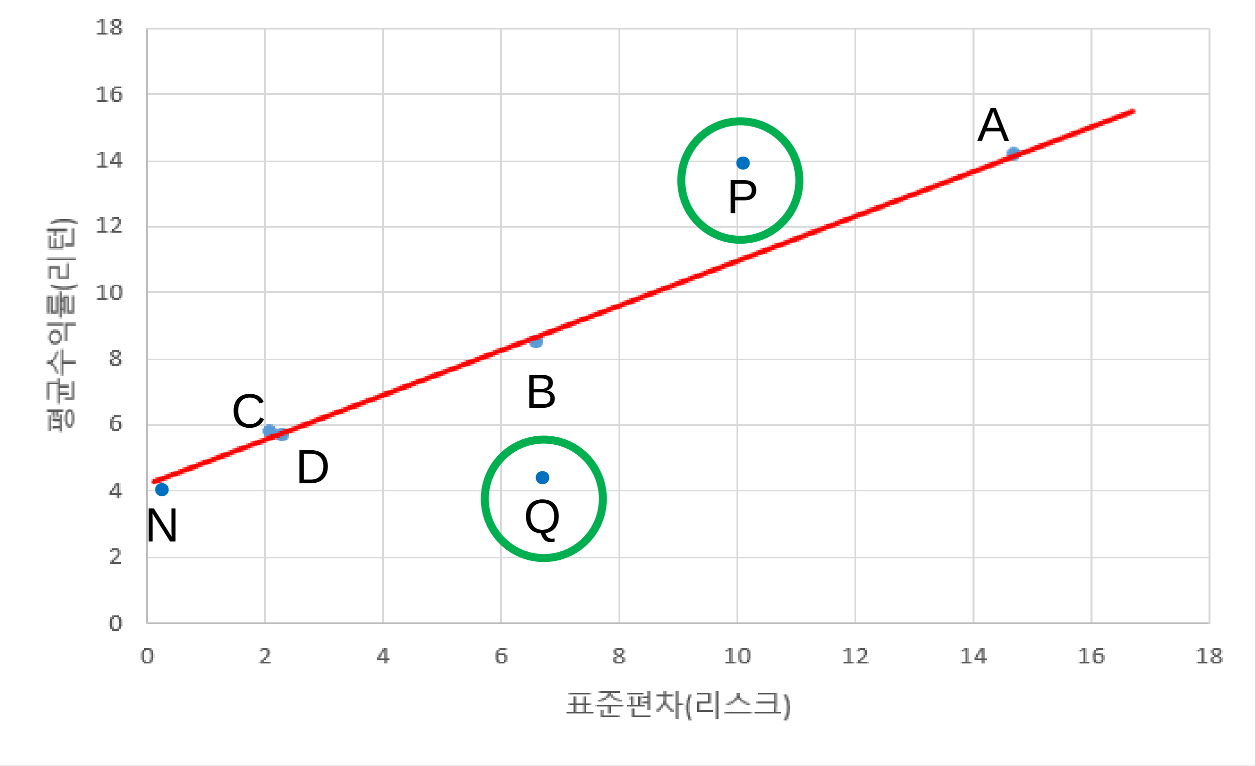
표준편차(3)(Cont.)

- **금융상품의 우열을 가리는 수치, 샤프지수**
  - 만일, 리턴이 30이고 표준편차가 3인 금융상품은 표준편차 1당 환산하면 30 ÷ 3 = 10의 리턴이 되고,
  - 리턴이 40이고 표준편차가 5인 금융상품은 표준편차 1당 환산하면 40 ÷ 5 = 8의 리턴이 있게 된다.
  - 즉, 리턴이나 리스크가 다른 상품을 통일시켜 비교할 수 있다.

표준편차(3)(Cont.)

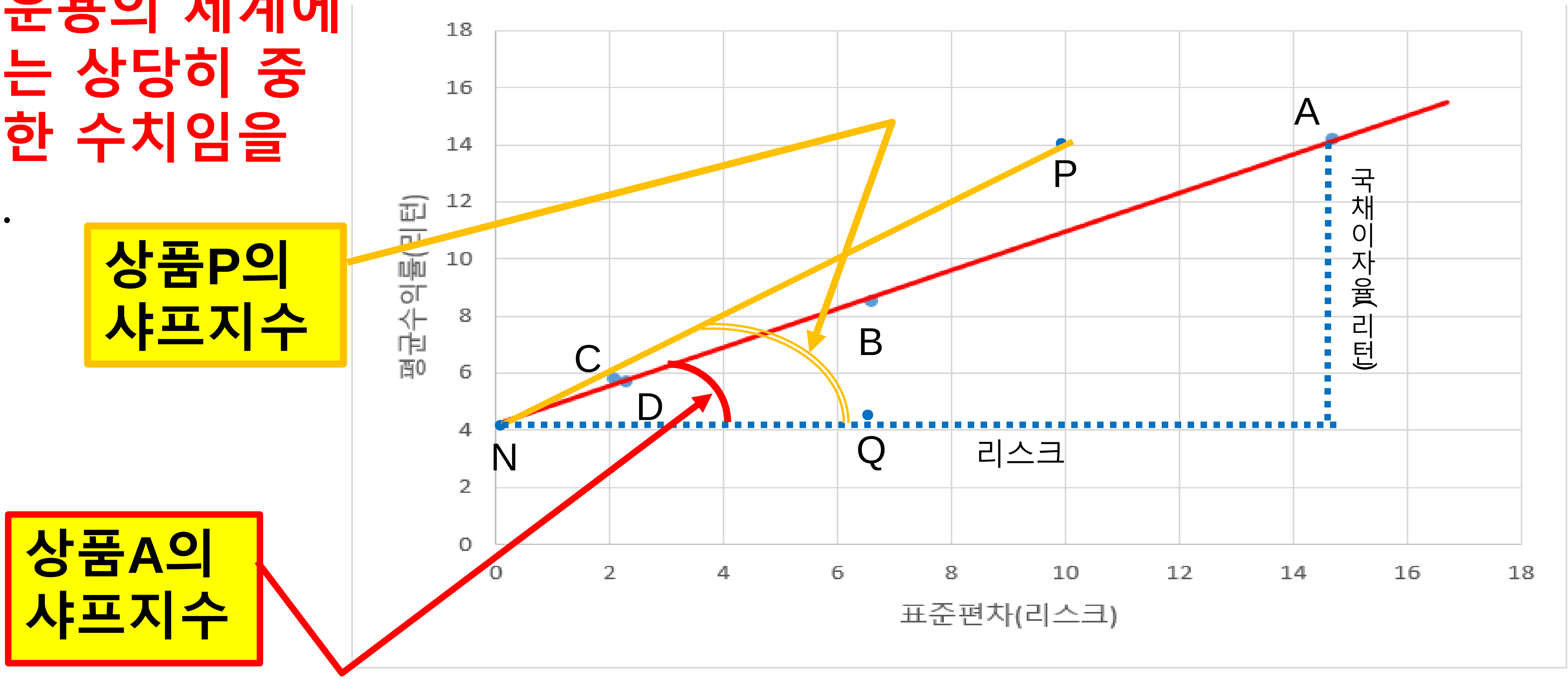
금융 상품의 우열을 가리는 방법

- 간단한게 현재 국채율이 4%라고 가정하자.
- 차트의 N점이다.
- 그러면 A상품의 샤프지수는 직선 NA의 기울기와 일치한다.
- 따라서, B, C, D상품의 기울기도 상품A의 기울기와 일치한다.
- 결론은, 샤프지수가 일치한다.
- 따라서, P상품은 A,B,C,D상품보다 운용을 잘하고, Q상품은 운영을 잘 못한다는 것을 알 수 있다.



표준편차(3)(Cont.)

- 금융 상품의 우열을 가리는 방법
  - 투자나 자산운용의 세계에서 표준편차는 상당히 중요하고 유효한 수치임을 알게 되었다.



표준편차(3)(Cont.)

- **금융상품의 우열을 가리는 수치, 샤프지수**
  - 다음은 생명보험회사의 운용실적을 나타낸 샤프지수이다.

	○○생명	☆☆생명	◎◎생명	◇◇생명	□□생명	△△생명	♤♤생명
평균	4	4.69	4.62	4.8	5.41	6.49	4.85
표준편차	5.48	4.47	5.59	4.28	5.64	4.64	6.43
샤프지수	0.107	0.286	0.216	0.324	0.354	0.663	0.223
순위	7	4	6	3	2	1	5

연습문제

- 운용실적이 평균수익률은 5%, 표준편차는 약 4.5%이다. 국채이자율이 3%라고 하면,

  샤프지수(SPM) = ( ) (소수점 둘째 자리)

- 샤프지수가 0.5인 투자신탁이 있었다고 가정해보자. 표준편차가 5%이고, 국채이자율이 3%라면, 이 투자신탁의 평균수익률은 ( )%이다.

정리

- 투자는 기본적으로 High Risk, High Return인 상품이나 Low Risk, Low Return인 상품 중에서 선택하게 된다. 이 상품의 차이는 성질의 차이이지, 우열을 의미하는 것은 아니다.
- 같은 평균수익률이라면 표준편차가 작은 것이 우량 금융상품이며, 같은 표준편차라면 평균수익률이 큰 것이 우량 상품이라고 할 수 있다.
- 이와 같은 의미에서, 금융상품의 우열을 평가하는 기준으로 샤프지수(SPM)이 있다. 이것은 (X의 샤프지수) = {(X의 리턴) – (국채 이율)} ÷ (X의 리스크)로 계산한다. 샤프지수가 큰 것이 우량 금융상품이라고 볼 수 있다.

정규분포

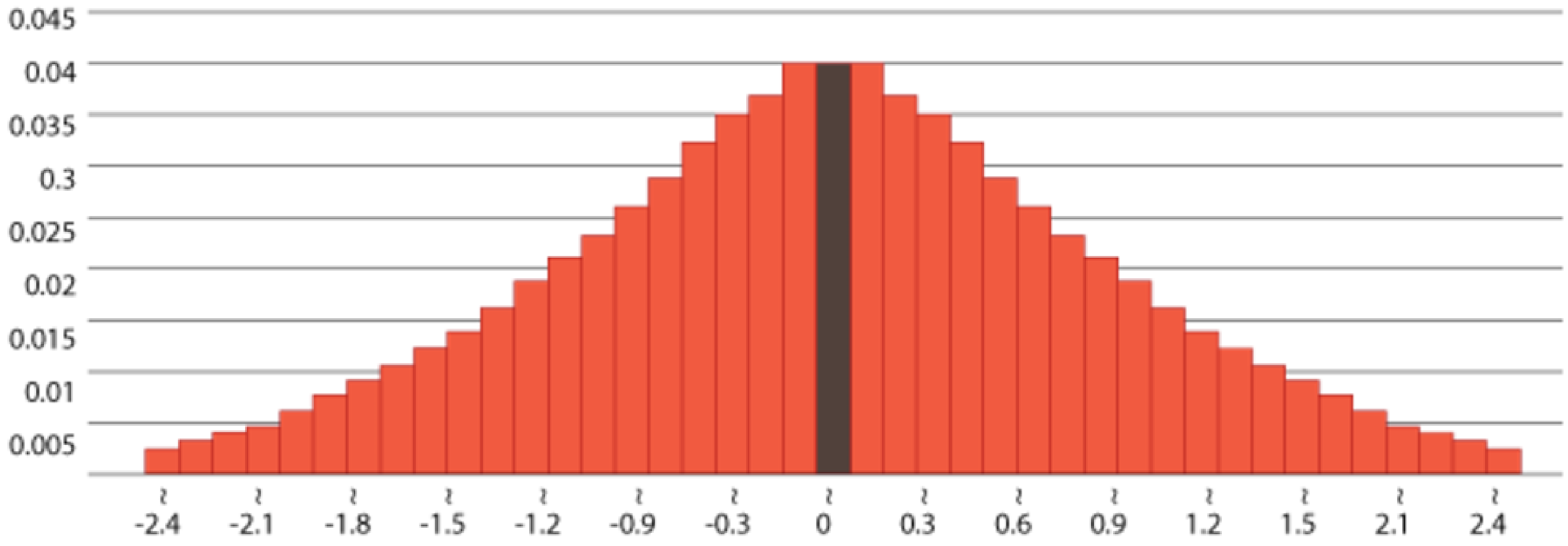
- **가장 많이 발견할 수 있는 데이터 분포**
  - 대부분의 데이터들은 그것들이 나타나는 불확실성의 구조를 반영한 것이다.
  - 대부분의 현상은 불확실성의 구조를 갖고 있으며, 발생하는 데이터는 제 각각의 값이 되는 경우가 일반적이다.
  - 데이터의 분포 : 데이터가 제 각각인 수치로 나타나는 것
  - 이제까지 데이터의 분포의 특징이나 반복되는 것을 파악하기 위한 도구로 평균값이나 표준편차라는 통계량을 설명했다.

정규분포 (Cont.)

- **가장 많이 발견할 수 있는 데이터 분포**
  - 데이터 분포에서 가장 대표적인 것은 ?
  - 이것은 자연이나 사회에서 관측되는 데이터들 속에 매우 자주 등장하는 것이다.
  - 동시에, 이런 분포의 모습은 수학적으로 정확히 설명되는 것이다.
  - 그것은 바로, 정규분포라고 하는 분포이다.
  - 사람이나 생물의 키 데이터, 주식의 수익률 데이터 등...

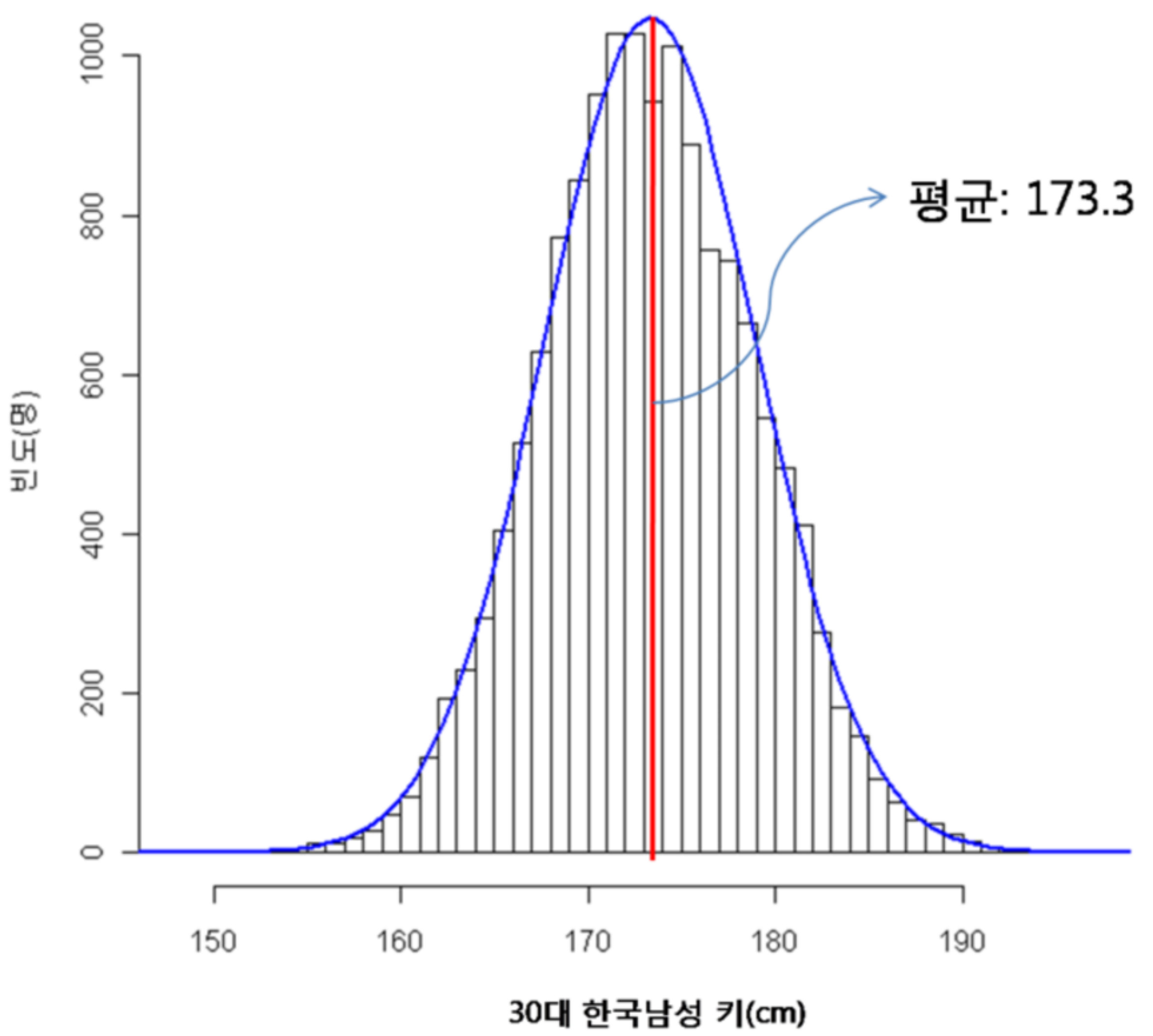
정규분포 (Cont.)

- **가장 많이 발견할 수 있는 데이터 분포**
  - 표준정규분포 : 정규분포 중에서 가장 기초
  - 표준정규분포 데이터 세트는 -∞에서 ∞까지 모든 수치의 데이터로 구성



정규분포 (Cont.)

- **가장 많이 발견할 수 있는 데이터 분포**
  - 표준정규분포 : 정규분포 중에서 가장 기초
  - 표준정규분포 데이터 세트는 -∞에서 ∞까지 모든 수치의 데이터로 구성



정규분포 (Cont.)

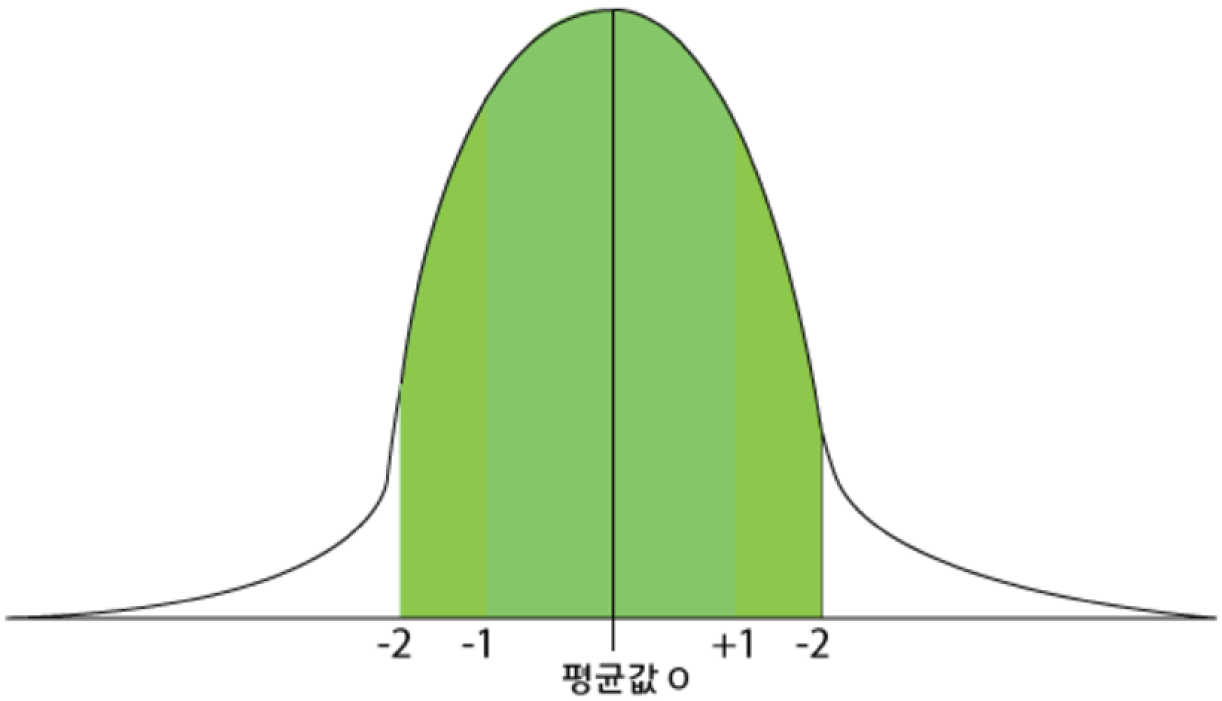
- **가장 많이 발견할 수 있는 데이터 분포**
  - 사실 부드러운 곡선인 그래프는 얇은 막대그래프를 모아 비슷하게 표현한 것이다.
  - 각 막대의 높이는 그 범위에 들어있는 무한개의 데이터가 많음을 나타내는 상대도수라고 가정한다.
  - 0 주변에 데이터가 집중해 있고(히스토그램이 높이가 높고), +2를 웃돌거나 -2를 밑돌면 데이터 수가 급격하게 줄어든다(히스토그램의 높이가 급격하게 낮아진다).

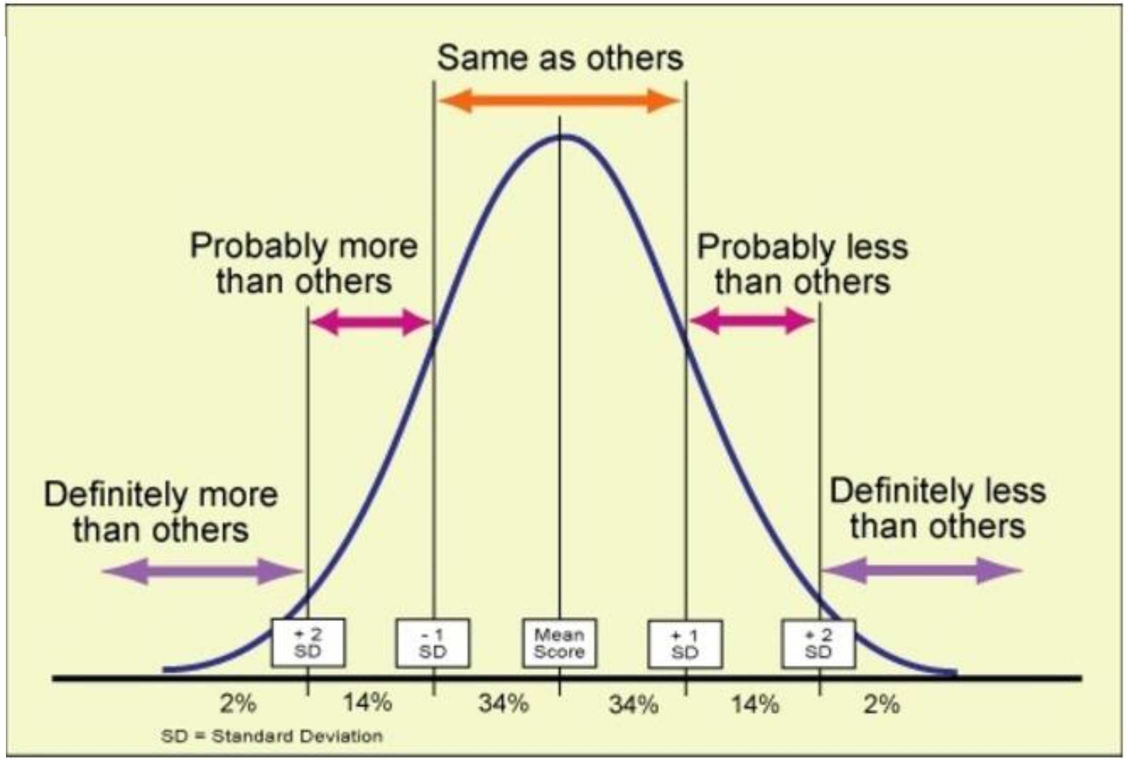
정규분포 (Cont.)

- **가장 많이 발견할 수 있는 데이터 분포**
  - 표준정규분포의 성질①
    - 평균값 = 0, 표준편차 = 1
    - 그래프가 0을 중심으로 좌우대칭이기 때문에 평균값은 0이다.
  - 표준정규분포의 성질②
    - (+1) ~ (-1) 범위의 데이터(평균에서 표준편차 1배 이내 범위의 데이터)의 상대도수는 0.6826 (=약 70%)
    - (+2) ~ (-2) 범위의 데이터(평균에서 표준편차 2배 이내 범위의 데이터)의 상대도수는 0.9544 (=약 95%)

정규분포 (Cont.)

- 가장 많이 발견할 수 있는 데이터 분포





정규분포 (Cont.)

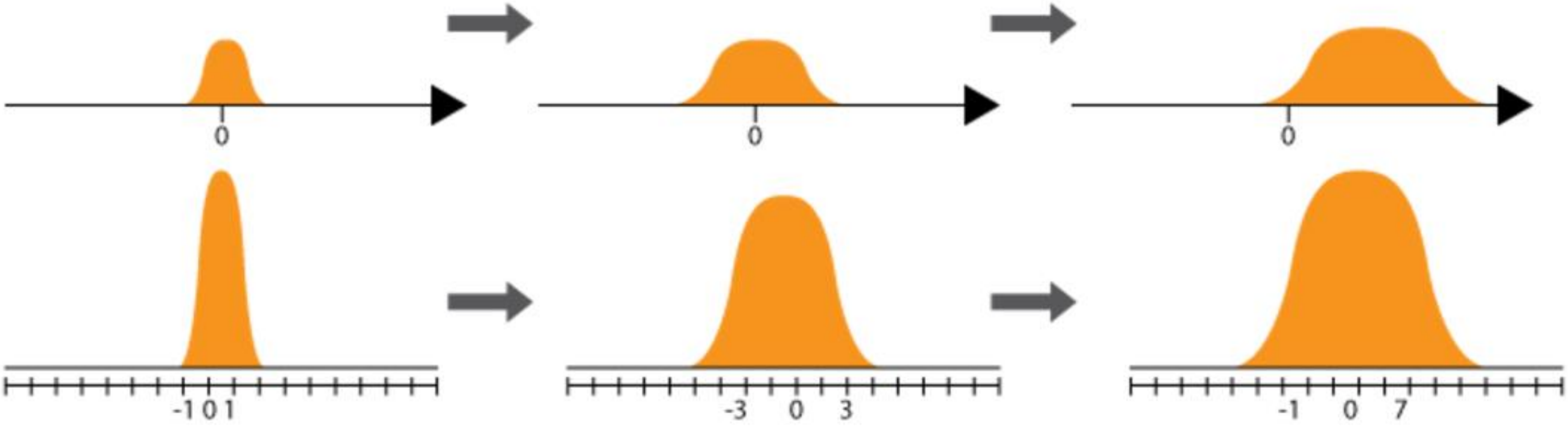
- **일반정규분포를 보는 방법**
  - 일반정규분포의 데이터 세트는 단순히 표준정규분포의 모든 데이터에 일정한 수를 곱하고, 그 뒤에 일정한 수를 더하는 방법으로 얻을 수 있다.
  - 곱하는 일정한 수를 σ(시그마), 더하는 일정한 수를 μ(뮤)라고 한다면,
  - **(일반정규분포의 데이터) = σ x (표준정규분포의 데이터) + μ**

정규분포 (Cont.)

- **일반정규분포를 보는 방법**
  - 일반정규분포의 성질①
    - σ x (표준정규분포의 데이터) + μ로 만들어진 데이터는
    - 평균값 = μ 표준편차 = σ
    - 만일, σ = 3이고, μ = 4라고 가정하자.
    - +1과 -1 사이에 있는 데이터의 상대도수는 대략 68%된다고 했기 때문에, 이것을 히스토그램ㅇ로 말하면 위로 블록한 그래프의 +1과 -1 사이에 있는 막대그래프가 전체에서 68%를 차지한다는 의미이다.
    - 그렇다면, 이 표준정규분포에서 데이터에 3을 곱하면 +3과 -3 사이에 있는 데이터의 상대도수는 대략 68%이고, 4를 더하면 +7과 +1 사이에 있는 데이터의 상대도수는 대략 68%라는 말이 된다.
    - 즉, 히스토그램은 좌우로 3배가 늘어나고 오른쪽으로 4만큼 이동한다.

정규분포 (Cont.)

- 일반정규분포를 보는 방법



정규분포 (Cont.)

- **일반정규분포를 보는 방법**
  - 일반정규분포의 성질②
    - (μ + 1 x σ) ~ (μ - 1 x σ)의 범위 데이터(평균에서 표준편차 1배 이내 범위의 데이터)의 상대도수는 0.6826 (=약 70%)
    - (μ + 2 x σ) ~ (μ - 2 x σ)의 범위 데이터(평균에서 표준편차 2배 이내 범위의 데이터)의 상대도수는 0.9544 (=약 95%)
  - 일반정규분포를 표준정규분포로 바꾸는 공식
    - 데이터 x가 평균값이 μ, 표준편차가 σ인 일반정규분포를 따르는 데이터일 경우, $z = (x - \mu) \div \sigma$라는 가공을 하면, 데이터 z는 표준정규분포를 따르는 데이터가 된다.

정규분포 (Cont.)

- 키 데이터는 정규분포를 따른다

계급	계급값	도수	상대도수	누적도수
141 – 145	143	1	0.0125	1
146 – 150	148	6	0.075	7
150 – 155	153	19	0.2375	26
156 – 160	158	30	0.375	56
161 – 165	163	18	0.225	74
166 – 170	168	6	0.075	80

여대생 80명 키의 ‘도수분포표 '

정규분포 (Cont.)

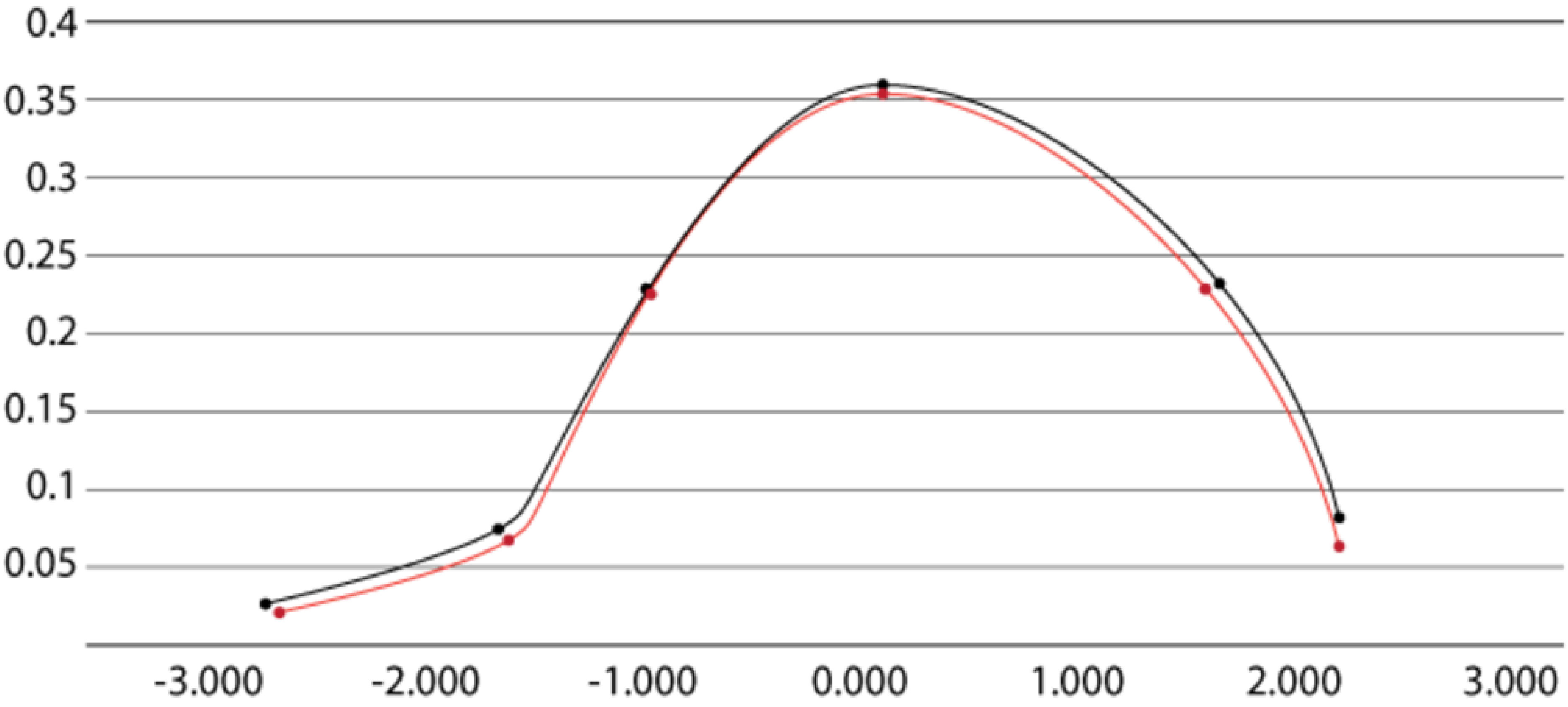
- **키 데이터는 정규분포를 따른다**

계급을 표준정규분포로 고친 값(z)	실제의 상대도수	정규분포로 가정할 때의 상대도수
-3.287 ~ -2.361	0.0125	0.0086
-2.361 ~ -1.435	0.075	0.0665
-1.435 ~ -0.509	0.2375	0.2297
-0.509 ~ 0.417	0.375	0.3563
0.417 ~ 1.343	0.225	0.2488
1.343 ~ 2.269	0.075	0.0781

평균값 = 157.75(cm), 표준편차 = 5.4

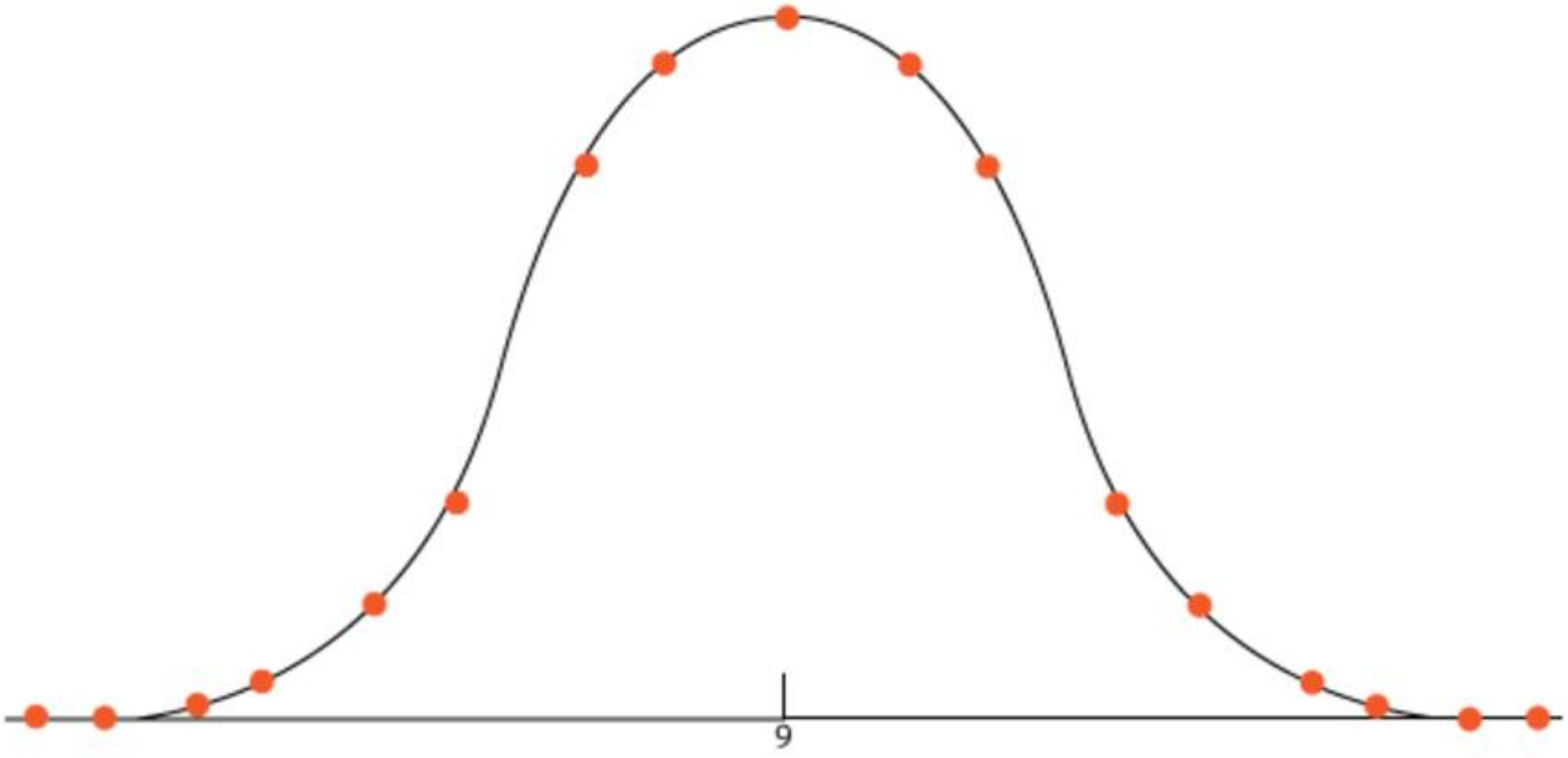
정규분포 (Cont.)

- 키 데이터는 정규분포를 따른다



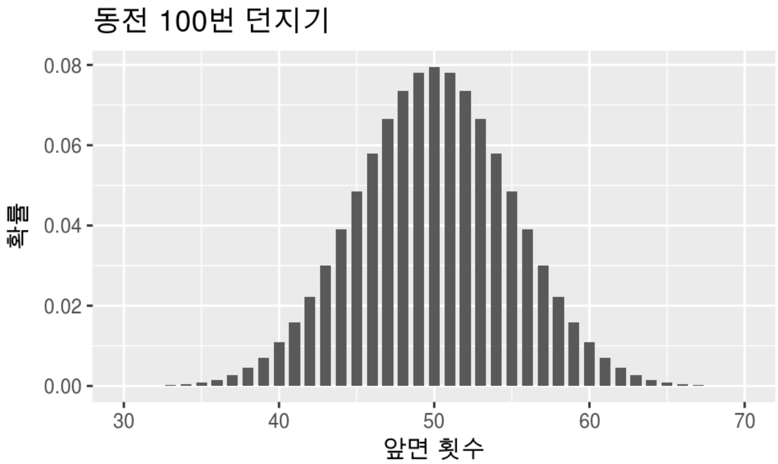
정규분포 (Cont.)

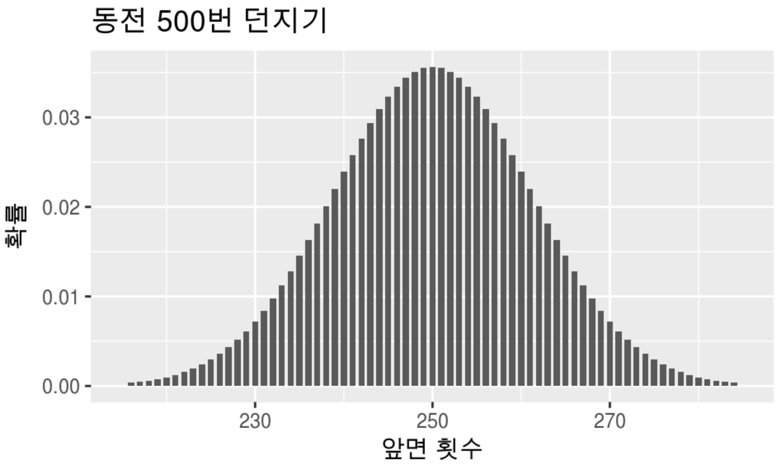
- **키 데이터는 정규분포를 따른다**
  - 동전을 18개 던졌을 때 앞면이 k개 나올 상대도수

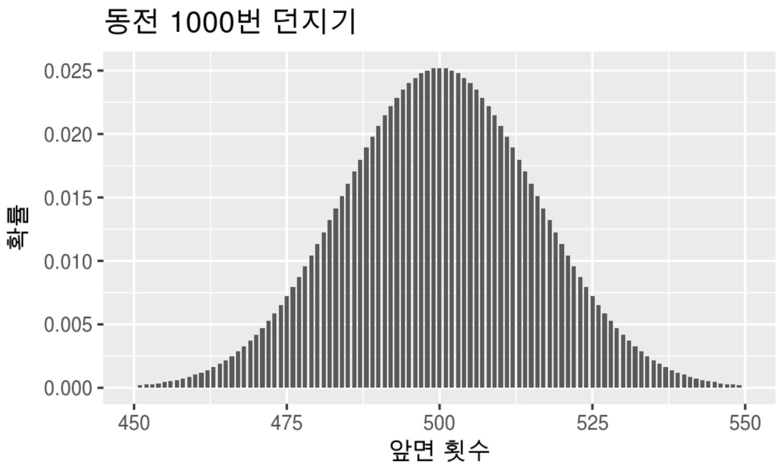


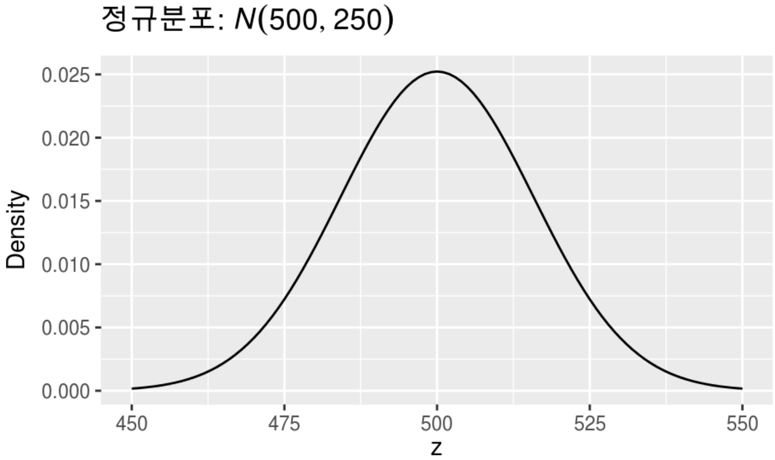
정규분포 (Cont.)

- **키 데이터는 정규분포를 따른다**
  - 동전을 N개 던졌을 때 나오는 표의 개수를 데이터로 나타냈을 때









정규분포 (Cont.)

- **키 데이터는 정규분포를 따른다**
  - 동전 던지기는 정규분포와 근사
    - 동전 N개를 동시에 던져서(혹은 N번 계속해서 던져서), 그 중 몇 개가(혹은 몇 번) 앞면으로 나올지를 데이터로 기록한다. 이 작업을 반복하여 앞면이 X 수가 나올 상대도수의 히스토그램을 만들면 그것은 근사적으로,
    - 평균값이 $\frac{N}{2}$, 표준편차가 $\frac{\sqrt{N}}{2}$ 인 정규분포를 따른다.

연습문제

- 1000점을 만점으로 하는 어떤 시험의 평균은 대략 600점이고, 표준편차가 100점이며, 정규분포를 한다고 한다. 이때, 95.44%의 데이터를 포함하는 범위는 (   ) – {(   ) X 2} ~ (   ) + {(   ) X 2}이기 때문에, (   ) ~ (   )의 범위가 된다.
- 100개의 동전을 동시에 던졌을 대 앞면이 나오는 동전의 수를 데이터로 집계하면 평균이 50개이고, 표준편차가 5개이며, 정규분포를 한다고 한다. 이때, 95.44%의 데이터를 포함하는 범위는 (   ) – {(   ) X 2} ~ (   ) + {(   ) X 2}이기 때문에, (   ) ~ (   )의 범위가 된다.

정리

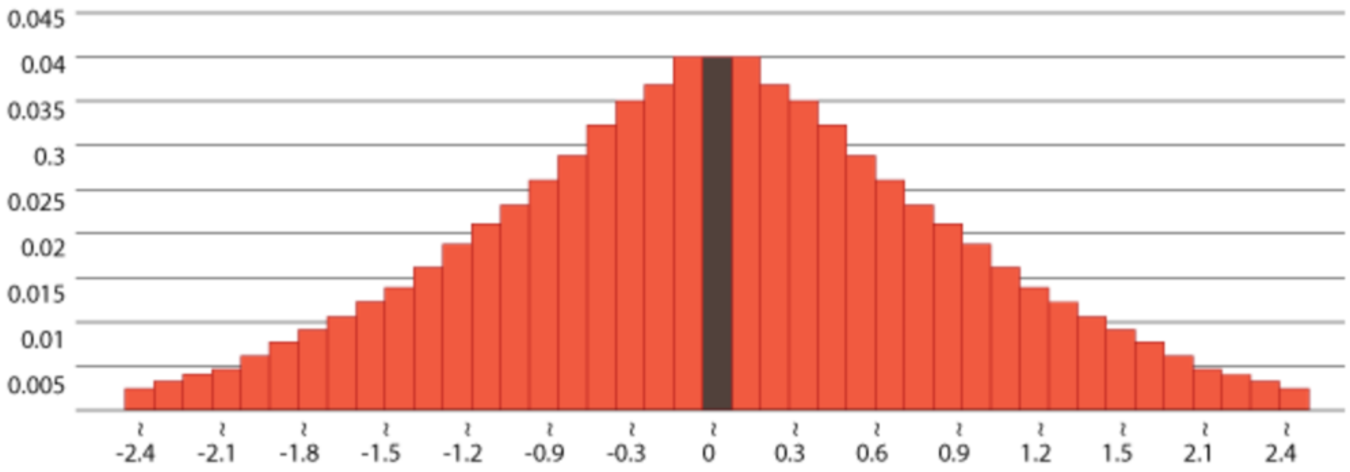
- 정규분포는 자연이나 사회에서 가장 흔히 볼 수 있는 분포다. 예를 들어, 키 데이터나 동전 던지기에서 앞면이 나올 개수의 데이터 등이 있다.
- 표준정규분포는 평균값 = 0이고, 표준편차 = 1 이다.
- 표준정규분포에서는 (+1) ~ (-1) 범위의 데이터(평균에서 표준편차 1배 이내의 범위에 있는 데이터)의 상대도수는 0.6826 (=약 70%), (+2) ~ (-2) 범위의 데이터(평균에서 표준편차 2배 이내의 범위에 있는 데이터)의 상대도수는 0.9544 (=약 95%)가 된다.

정리

- 일반정규분포의 데이터는 **σ x (표준정규분포의 데이터) + μ** 로 구하고, **평균값 = μ이고, 표준편차 = σ이다.**
- 평균값이 μ이고, 표준편차가 σ인 정규분포를 표준정규분포로 다시 구하기 위해서는 **z = (x – μ) ÷ σ** 라는 식을 적용하면 된다.
- 평균값이 μ이고, 표준편차가 σ인 정규분포에서는 (μ + 1 x σ) ~ (μ - 1 x σ)의 범위 데이터(**평균에서 표준편차 1배 이내 범위의 데이터)의 상대도수는 0.6826 (=약 70%),** (μ + 2 x σ) ~ (μ - 2 x σ)의 범위 데이터(**평균에서 표준편차 2배 이내 범위의 데이터)의 상대도수는 0.9544 (=약 95%)**이다.

통계적 추정의 출발점

- **정규분포의 성질을 이용해 예언을 할 수 있다**
  - 만일 주목하고 있는 불확실성 현상이 정규분포라고 간주한다면, 정규분포의 성질을 이용해서 어떠한 예언(예측)을 할 수 있지 않을까?
  - 그것이 바로 **통계적 추정**의 출발점이다.
  - 그 다음에 어떤 데이터가 발생할지 모르지만, 그 상대도수는 표준정규분포를 따른다는 사실.



통계적 추정의 출발점 (Cont.)

- **정규분포의 성질을 이용해 예언을 할 수 있다**
  - 예측을 맞추기 위해서는 → 나타날 가능성이 큰 수
  - 히스토그램의 막대 높이는 데이터가 나타나는 상대도수이므로, 이것은 나타날 가능성을 보여주는 것
  - 따라서 앞 슬라이드의 히스토그램에서 막대의 높이가 높은 것은 0에 가깝다.
  - 그러므로, **0에 가까운 것을 예측하는 것이 쉽게 맞추기 위한 좋은 전략**이다.
  - 예측을 하기 위해서 폭을 지정해서 O이상 O이하라는 식으로 하는 것이 좋다.

통계적 추정의 출발점 (Cont.)

- **정규분포의 성질을 이용해 예언을 할 수 있다**
  - 예측을 하기 위해서 폭을 지정해서 O이상 O이하라는 식으로 하는 것이 좋다.
  - 즉, 0이상 0.1이하의 수라고 예측하면, 이 구간은 데이터 상대도수는 약 0.04이다. → 표준정규분포 데이터의 약 4%는 이 구간의 수치라고 할 수 있기 때문에, 0이상 0.1이하의 수라고 예측하면 맞출 확률은 4%라고 해도 좋다.
  - 그러면 예측의 정확성을 만족시킬 수준까지 높이려면?
  - 범위가 -1에서 +1까지 데이터의 상대도수는 약 68.26%이므로, 약 68.26%의 확률로 그 예측을 맞출 수 있다는 말이다. → 적중률이 상당히 높아졌다.

통계적 추정의 출발점 (Cont.)

- **표준정규분포의 95% 예측적중구간**
  - 적중확률을 높이고 싶으면 구간을 넓혀야 한다.
  - 현실적으로 -∞에서 +∞이하의 수로 예측한다는 것이 불가능하기 때문에, 유한한 범위에서 할 수 밖에 없다.
  - 많이 사용되는 것은 95% 적중 또는 99% 적중의 범위이다.
  - 만일 95% 적중이 범위를 고른다는 말은 5%의 예측은 틀린다라는 말이다.
  - 통계학에서는 적중확률을 가능한 한 95%로 고정한다.
  - 그러면, -2이상 +2이하의 수의 상대도수가 약 95.44%이기 때문에 남은 0.44만큼을 삭제하기 위해 구간을 **-1.96이상 +1.96이하**로.

통계적 추정의 출발점 (Cont.)

- **표준정규분포의 95% 예측적중구간**
  - 표준정규분포의 95% 예측적중구간은 -1.96 이상 +1.96 이하이다.
  - 처음부터 100% 맞추는 것은 불가능하다는 전제
  - 이것은 예측적중구간의 개념은 5%는 틀린다는, 완벽하지 않다는 점을 허용하는 것으로, 상당히 좁은 구간의 예측을 가능하게 하는 것이라고 이해한다.
  - 예측의 정확성만으로 보면, 예측하는 구간은 짧으면 짧을 수록 좋다.
  - 같은 예측적중 확률의 구간 중에서 가장 짧은 구간을 선택하는 길은 **좌우대칭의 구간**을 선택하는 것이다.

통계적 추정의 출발점 (Cont.)

- **일반정규분포의 95% 예측적중구간**
  - (일반정규분포의 데이터) = σ x (표준정규분포의 데이터) + μ
  - 일반정규분포의 95% 예측적중구간
    - 평균값이 μ이고, 표준편차가 σ인 정규분포의 95% 예측적중구간은 (μ – 1.96 * σ) 이상 ((μ + 1.96 * σ)이하 이다.
  - 일반정규분포를 표준정규분포로 바꾸는 공식
    - 데이터 x가 평균값이 μ이고, 표준편차가 σ인 일반정규분포를 따르는 데이터일 때, z = (x – μ) ÷ σ라는 가공을 하면, 데이터 z는 표준정규분포를 따르는 데이터가 된다.

통계적 추정의 출발점 (Cont.)

- **일반정규분포의 95% 예측적중구간**
  - 일반정규분포의 95% 예측적중구간 : 부등식 표시
    - 데이터가 x가 평균값이 μ이고, 표준편차가 σ이며, 일반정규분포를 따르는 경우일 때, 95% 예측적중구간은 부등식 $-1.96 \le \frac{x-\mu}{\sigma} \le +1.96$을 풀어서 구한 범위이다.
  - 예를 들어, N개의 동전을 던져서 앞면이 나오는 개수는 대략 평균값이 $\frac{N}{2}$, 표준편차가 $\frac{\sqrt{N}}{2}$ 인 일반정규분포가 된다. 그래서 100개의 동전을 동시에 던질 때 앞면이 나오는 개수를 몇 번 반복해서 관찰 후 상대도수 히스토그램을 만들면, 평균값이 $\frac{100}{2} = 50$이고, 표준편차가 $\frac{\sqrt{100}}{2} = 5$ 인 일반정규분포를 하는 히스토그램과 닮는다.

통계적 추정의 출발점 (Cont.)

- **일반정규분포의 95% 예측적중구간**
  - 다시 말해서, 지금부터 100개의 동전을 동시에 던졌다고 가정해보자. 앞면이 나올 개수를 예측한다면, 95% 예측 적중할 범위를 만들 수 있다.
  - (μ – 1.96 * σ) 이상 ((μ + 1.96 * σ)이하를 예측하면 좋기 때문에, μ = 50, σ = 5를 대입하면, (50 – 1.96 x 5)이상 (50 + 1.96 x 5) 이하 = 40.2 이상 59.8 이하가 95% 예측적중 범위가 된다.
  - 즉, 앞면이 나오는 동전은 40개에서 60개 사이라고 예측하면 대략 맞는다.
  - 대략의 의미는 충분히 많은 횟수인 M번을 예측하면 그 중 5%의 횟수(M x 0.05번)는 예측이 틀리다는 의미이다.

통계적 추정의 출발점 (Cont.)

- **일반정규분포의 95% 예측적중구간**
  - 부등식 표시
    - -1.96 ≤ $\frac{x-\mu}{\sigma}$ ≤ +1.96에서 μ에 50, σ에 5를 대입하면,
    - -1.96 ≤ $\frac{x-50}{5}$ ≤ +1.96에서 세 곳에 5배를 하면,
    - -1.96 x 5 ≤ $\frac{x-50}{5}$ x 5 ≤ +1.96 x 5
    - -9.8 + 50 ≤ x – 50 ≤ + 9.8에서 세 곳에 50을 더하면,
    - -9.8 + 50 ≤ x – 50 + 50 ≤ +9.8 + 50
    - 40.2 ≤ x ≤ 59.8

연습문제

- 여성의 키 평균값은 약 160cm, 표준편차는 약 10cm인 정규분포라고 알려져 있다. 당신이 내일 만날 여성의 키를 예측한다면 했을 때, 95% 적중시키려면 어느 범위를 예측하면 좋을까?
- 부등식

$$-1.96 \leq \frac{x-(\qquad)}{(\qquad)} \leq +1.96$$ 을 풀고

(　　　　　　　)cm 이상 (　　　　　　　　　　)cm 이하라고 예측하면 된다.

정리

- 표준정규분포의 95% 예측적중구간은 -1.96 이상 +1.96이하다.
- 평균값이 μ이고, 표준편차가 σ인 정규분포의 95% 예측적중구간은 (μ – 1.96 x σ) 이상 (μ + 1.96 x σ) 이하이다.
- 데이터 x가 평균값이 μ이고, 표준편차가 σ인 일반정규분포를 따르는 데이터일 때, z = (x – μ) ÷ σ라는 계산을 하면, 데이터 z는 표준정규분포를 따르는 데이터가 된다.
- 데이터 x의 평균값이 μ이고, 표준편차가 σ인 정규분포를 따를 경우, 95% 예측적중구간은 부등식 $-1.96 \le \frac{x-\mu}{\sigma} \le +1.96$ 을 풀어서 구한 범위이다.