기술통계 : 표와 그래프
Descriptive Statistics: Tabular and Graphical Methods

기술통계
(Descriptive Statistics)

- 표(Tables)
- 그래프나 수치(Graphs or figures)
- 통계치(Statistic)

빈도분포 (Frequency Distribution)

- 중복되지 않는 여러 집단의 각각의 항목들에 대한 빈도(혹은 숫자)를 표로 요약한 자료.
- 원본 자료를 통해 바로 얻을 수 없는 정보를 제공하는 것이 목표.

Example: Guest House

- 에리카 캠퍼스 게스트 하우스에 머무는 고객에게 숙박 시설의 품질에 대한 점수 평가하도록 함.
  - Excellent (우수),
  - Above average (평균이상),
  - Average (평균),
  - Below average (평균이하),
  - Poor (부족)
- 20명의 고객으로부터 제공된 점수는 다음과 같다.

Evaluation on Customer Satisfaction of Guest House

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	

빈도분포(Frequency Distribution)

Rating	Frequency
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
Total	20

빈도 분포 (Frequency Distribution)

- 집단의 개수를 정하기 위한 가이드라인
  - 5-20개 사이의 집단을 사용하라.
  - 일반적으로 자료의 수가 많을 수록 많은 수의 집단을 필요로 한다.
  - 반대로, 자료의 수가 적을 수록 적은 수의 집단을 필요로 한다.

빈도 분포
(Frequency Distribution)

- 집단의 범위를 정하기 위한 가이드라인
  - 동일한 범위의 집단들을 사용하라.
  - 집단의 범위는 대략 다음과 같이 구할 수 있다.

$$\frac{\text{가장 큰 값} - \text{가장 적은 값}}{\text{집단의 개수}}$$

Example:
현대자동차 서비스

- 현대 자동차 서비스의 매니저는 엔진의 튜업(tune-up) 부품 경비 리스트를 보다 보기 좋게 만들고자 함.
- 50명의 고객 송장 샘플을 통해 얻은 부품 경비는 다음과 같다.

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

Example: 현대자동차 서비스

- 빈도 분포 6개의 집단을 선택할 경우

  대략적인 집단의 범위 = (109 - 52)/6 = 9.5 ≒ 10

Cost ($)	Frequency
50-59	2
60-69	13
70-79	16
80-89	7
90-99	7
100-109	5
Total	50

상대빈도 분포 (Relative Frequency Distribution)

- 한 집단의 상대빈도(relative frequency)는 그 집단의 빈도가 총 도수(total number of data)에 대해 가지는 비율.

- 상대빈도분포(relative frequency distribution)는 각 집단의 상대빈도를 나타내는 자료를 도표로 요약한 것.

백분율빈도 분포 (Percent Frequency Distribution)

- 한 집단의 백분율빈도(percent frequency)는 상대빈도에 100을 곱한 값.
- 백분율빈도분포(percent frequency distribution)은 각 집단의 백분율을 나타내는 자료를 도표로 요약한 것.

Example: 게스트 하우스

- 상대도수와 백분율빈도 분포

Rating	Relative Frequency	Percent Frequency
Poor	.10	10
Below Average	.15	15
Average	.25	25
Above Average	.45	45
Excellent	.05	5
Total	1.00	100

Example: 현대 자동차 서비스

- 상대도수와 백분율빈도 분포

Cost ($)	Relative Frequency	Percent Frequency
50-59	.04	4
60-69	.26	26
70-79	.32	32
80-89	.14	14
90-99	.14	14
100-109	.10	10
Total	1.00	100

Example: 현대 자동차 서비스

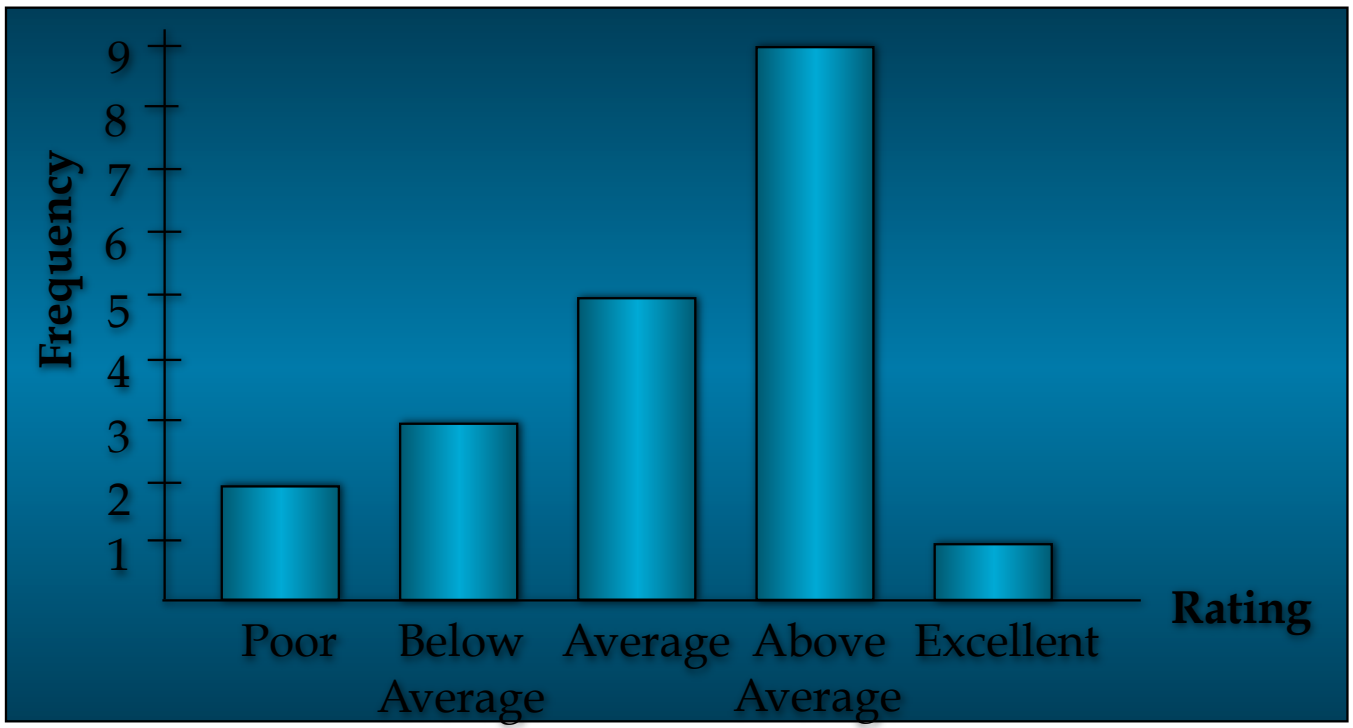
- 백분율빈도분포를 통해 얻은 정보
  - 부품비용 $50-59집단은 전체의 4%에 불과함.
  - 부품비용 $70이하는 전체의 30%임.
  - 가장 큰 비율(32%, 거의1/3)을 차지한 집단은 부품비용 $70-79 집단임.
  - 부품비용 $100 혹은 그 이상인 집단은 전체의 10%임.

막대그래프(Bar Graph)

- 막대 그래프(bar graph)는 정성적 자료(qualitative data)를 나타내는 그래픽 도구.
- 수평축에는 각 집단의 라벨을 표시함.
- 빈도, 상대빈도, 백분율은 수직 축에 표시함.
- 각 집단의 라벨 위에 그려진 고정폭 막대(bar of fixed width)를 사용해서 높이를 적절하게 조정.
- 분리된 막대(bars are separated)를 통해 각각의 독립된 집단을 분석.

Example: 게스트 하우스

- 막대그래프(Bar Graph)

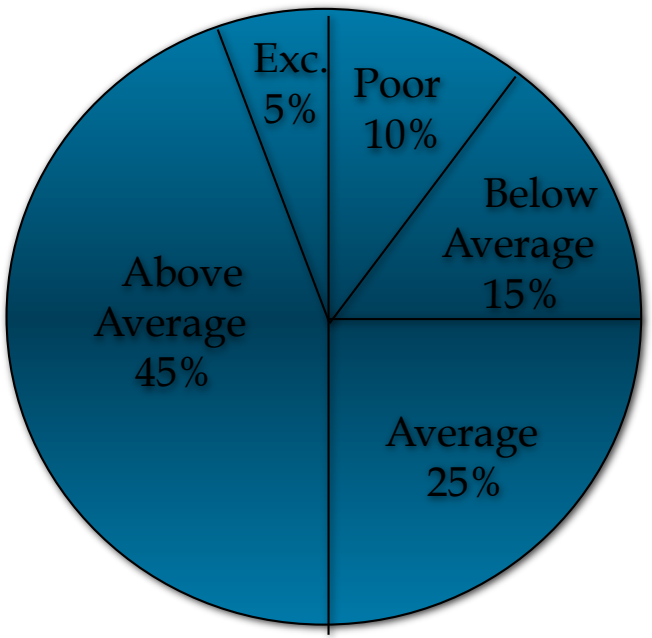


파이 차트(Pie Chart)

- 파이차트(Pie Chart)는 일반적으로 정성적 자료를 상대빈도분포로 나타내기 위한 그래픽 도구로 사용됨.
- 원을 그린 후, 각 집단의 상대빈도에 부합하는 부채꼴(sector)의 형태로 원을 나눔.
- 하나의 원은 360도로 이루어져 있으므로, 상대빈도 .25의 집단은 .25(360) = 90도 만큼의 부채꼴로 나타낼 수 있음.

Example: 게스트 하우스

- 파이 차트(Pie Chart)



Quality Ratings

Example: 게스트 하우스

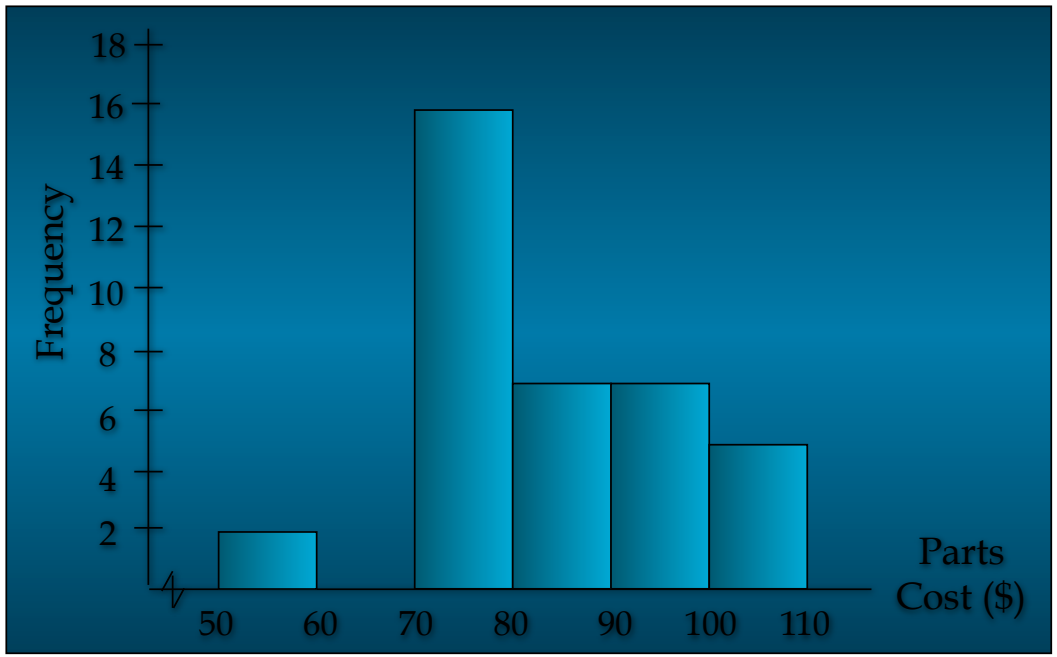
- 파이차트(Pie Chart)를 통해 다음과 같은 정보를 얻을 수 있음.
  - 설문에 응답한 고객 중 절반(1/2)이 프레지던트 호텔의 품질에 대해 'above average(평균이상)' 혹은 'excellent(우수)'라고 평가함.
  - 한 명의 고객이 'excellent(우수)'라고 평가하였으며, 두 명의 고객이 'poor(부족)'이라고 평가.

히스토그램(Histogram)

- 히스토그램은 정략적 자료(quantitative)를 그래픽으로 나타내는 또 하나의 일반적인 방식.
- 측정하고자 하는 변수를 수평축에 표시함.
- 각 집단 별로 그 구간의 빈도, 상대빈도, 백분율에 부합하는 사각형을 그림.
- 막대 그래프(bar graph)와 달리, 히스토그램은 인접 집단의 사각형과 분리되어있지 않음.

Example: 현대자동차 서비스

- 히스토그램(Histogram)



누적분포 (Cumulative Distributions)

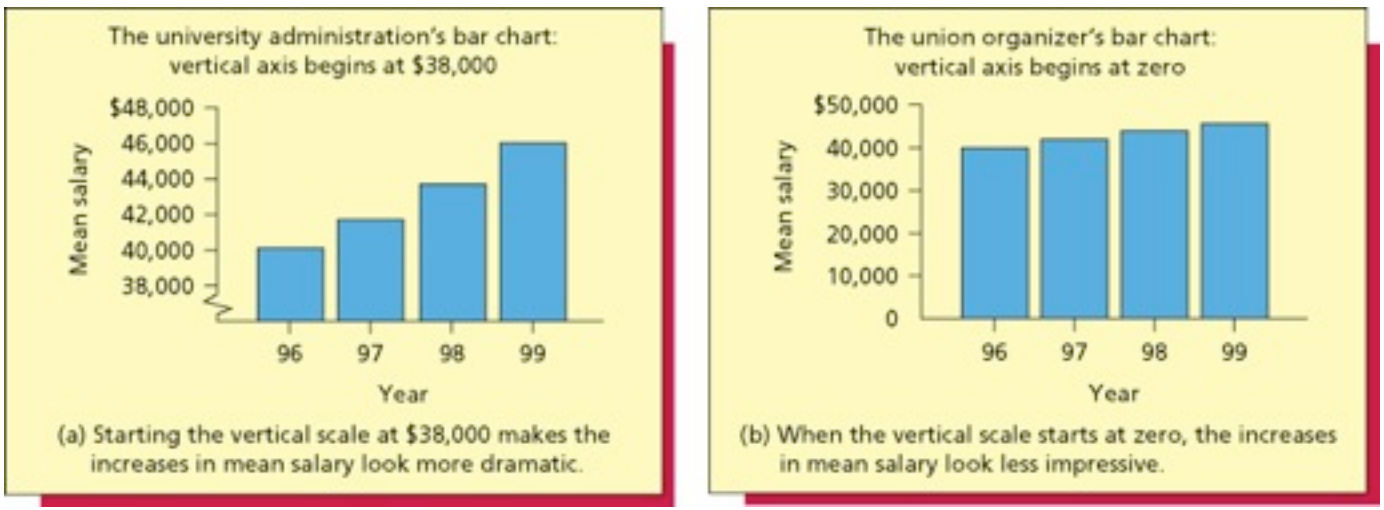
- 누적빈도분포(Cumulative frequency distribution)
  - 각 집단의 상한 점과 동일하거나 그보다 낮은 값들의 빈도를 나타냄.
- 누적상대빈도분포 (Cumulative relative frequency distribution)
  - 각 집단의 상한 점과 동일하거나 그보다 낮은 값들의 비율을 나타냄.

Example: 현대자동차 서비스

- 누적분포(Cumulative Distributions)

Cost ($)	Cumulative Cumulative Frequency	Relative Frequency
≤ 59	2	.04
≤ 69	15	.30
≤ 79	31	.62
≤ 89	38	.76
≤ 99	45	.90
≤ 109	50	1.00

잘못된 그래프와 차트 : Scale Break

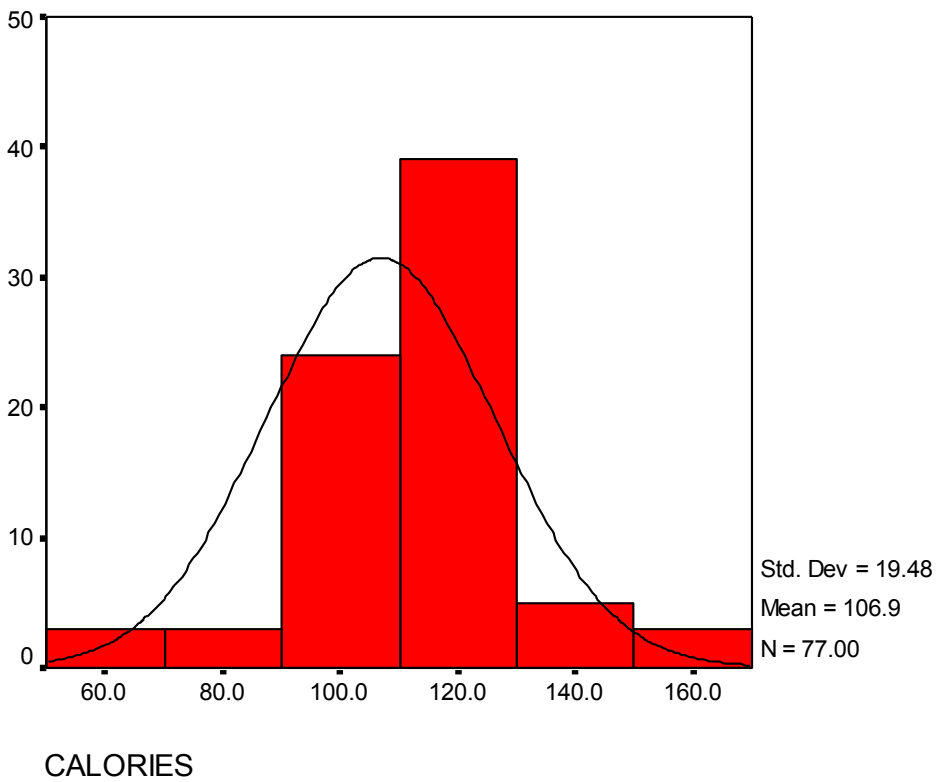


Mean Salaries at a Major University, 1996 - 1999

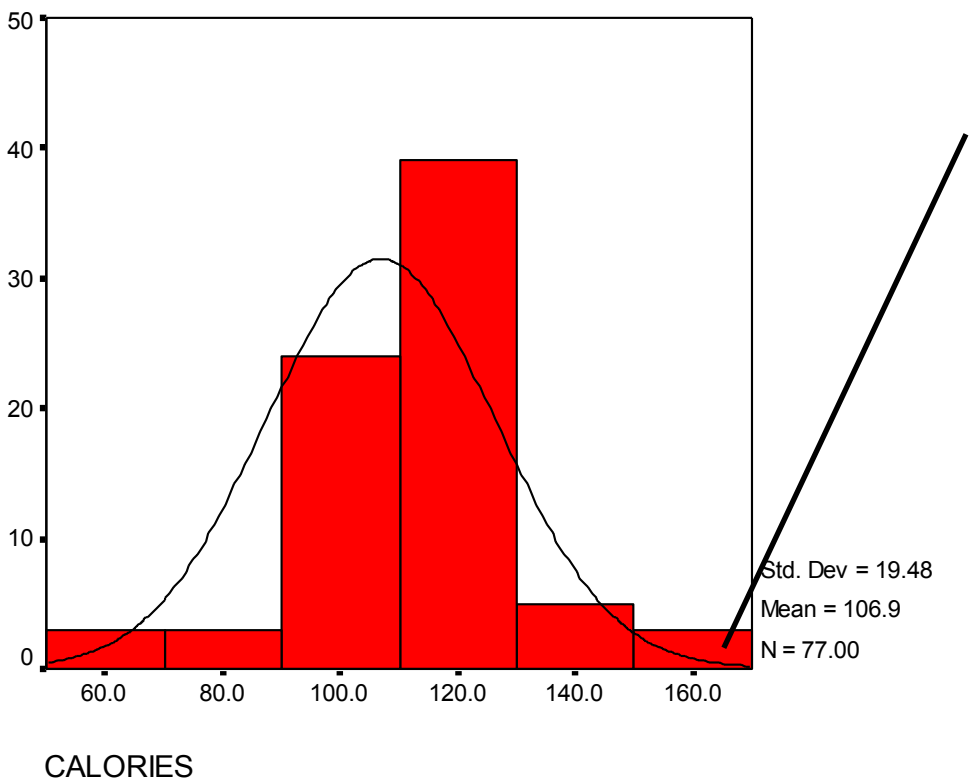
히스토그램의 형태 (Histogram shapes)

- 히스토그램은 단봉(unimodal, 하나의 최빈값), 이봉(bimodal, 두 개의 최빈값), 아니면 다봉(multimodal, 두 개 이상의 최빈값)을 가질 수 있음.
- 히스토그램은 U형태(U-shaped)분포와 종모양(Bell-shaped)분포를 가질 수 있음.
- 히스토그램은 종종(항상은 아님) 좌우대칭를 이루고, 좌우대칭이 아닐 경우 비대칭를 이룸.

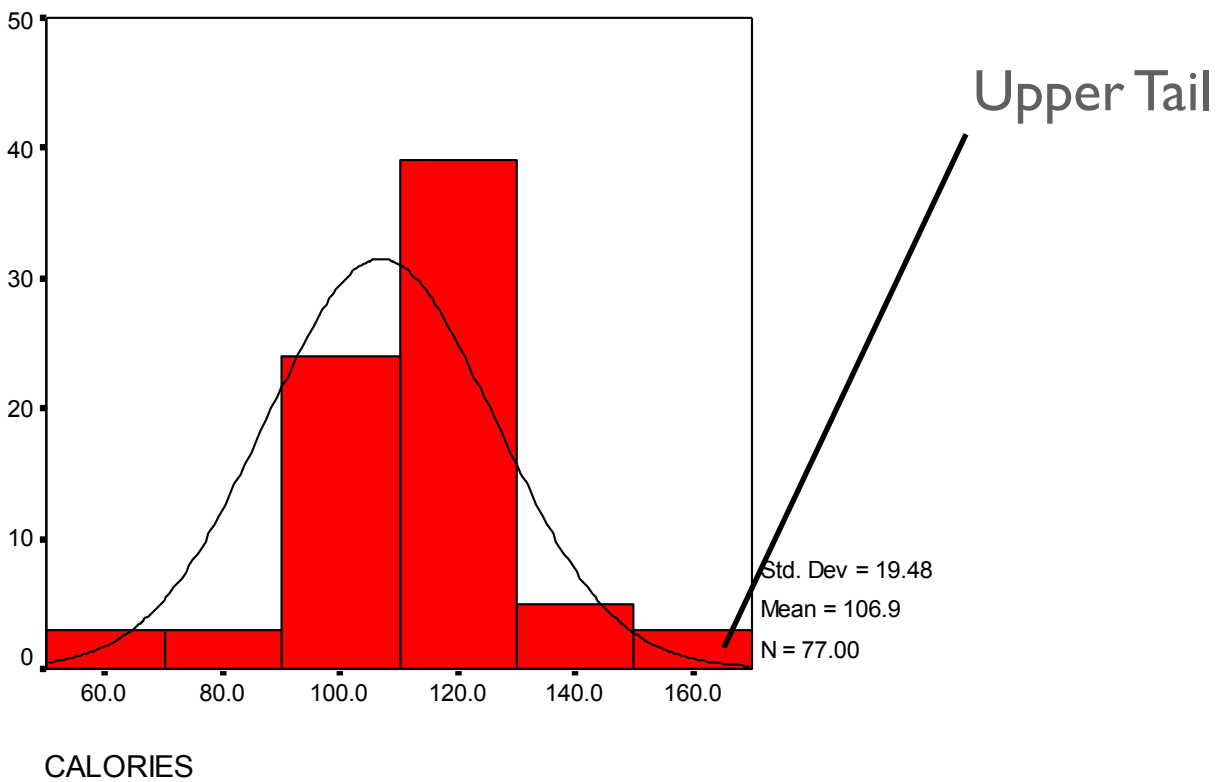
Sample histogram



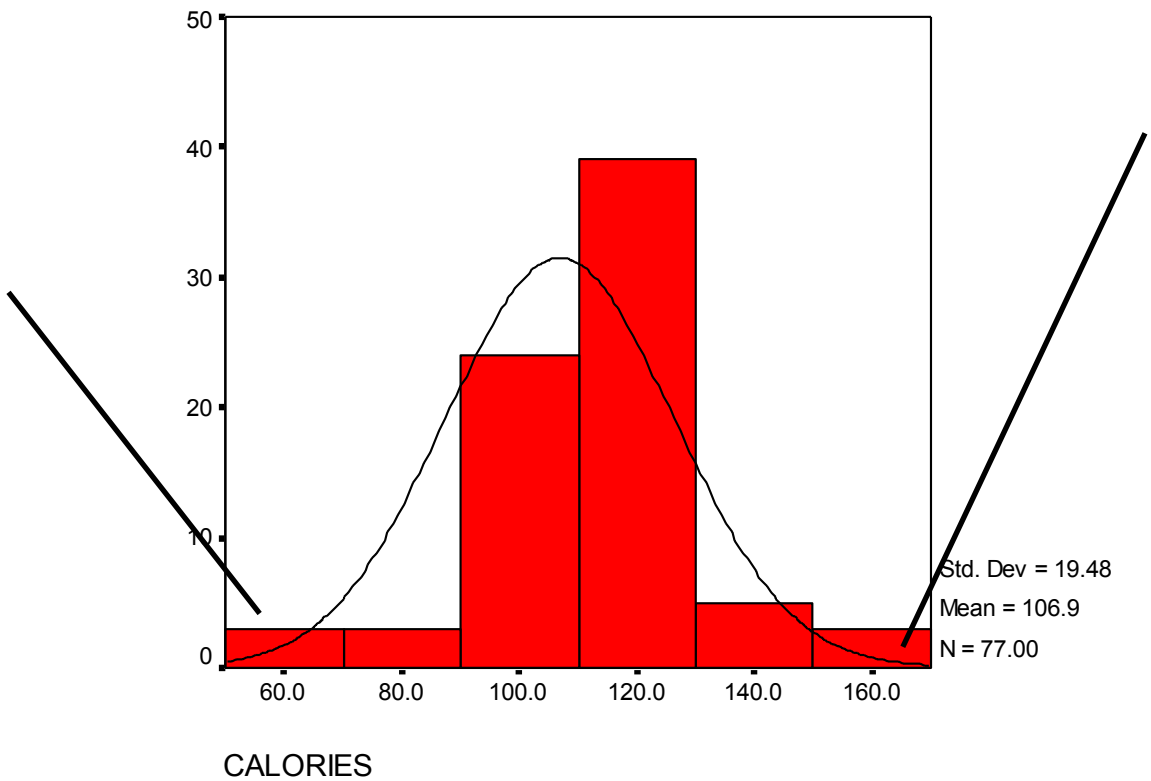
Sample histogram



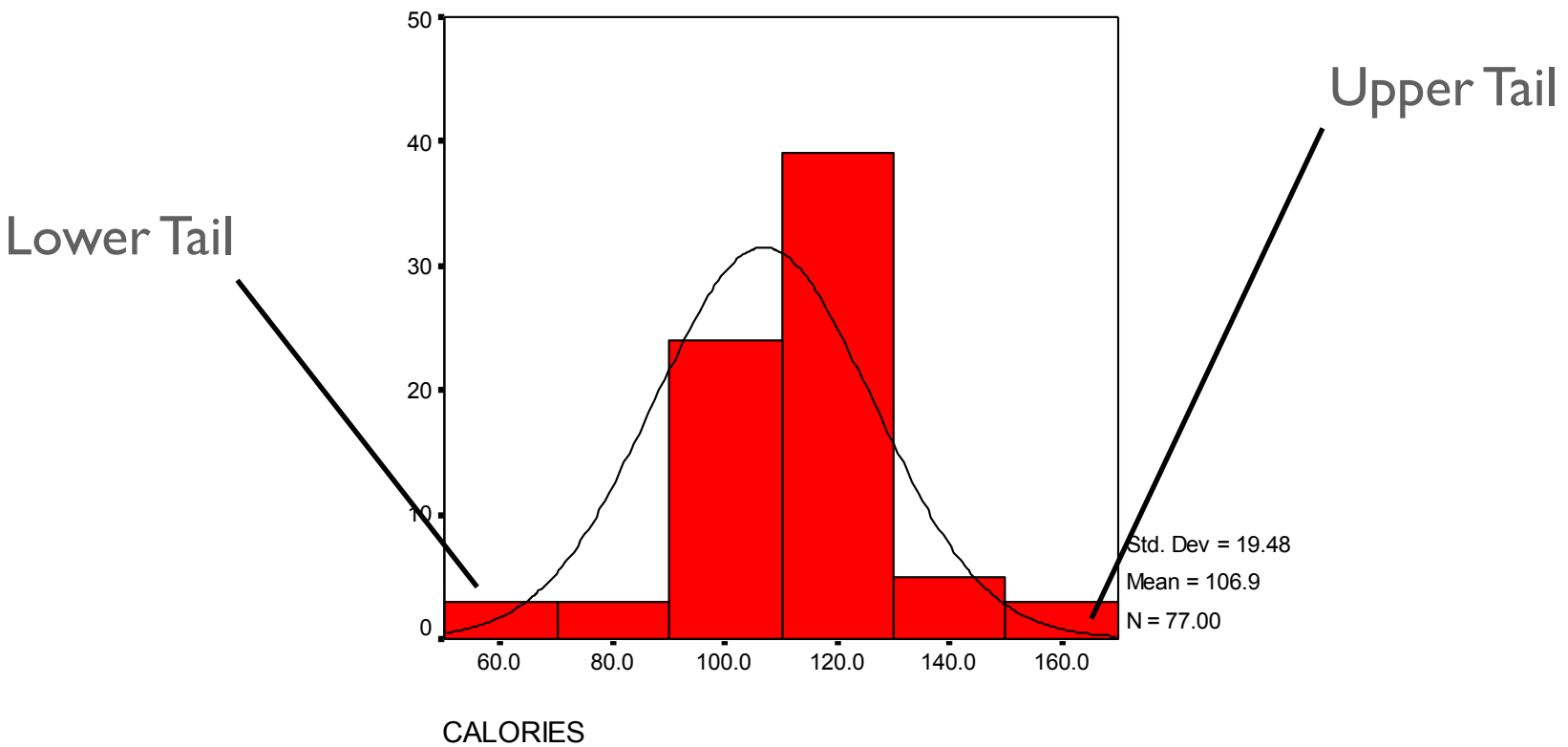
Sample histogram



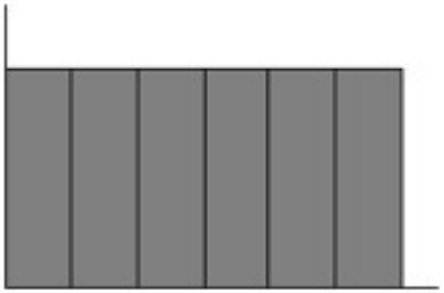
Sample histogram



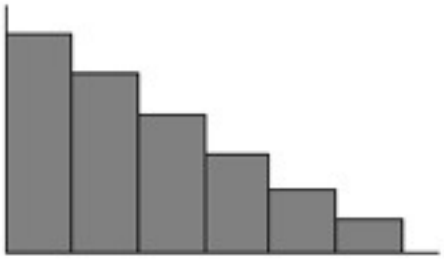
Sample histogram



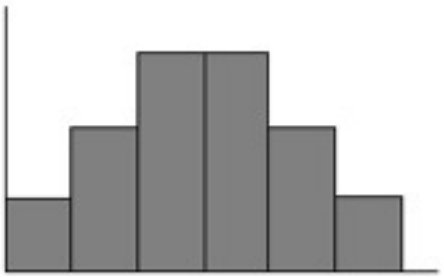
Histogram shapes



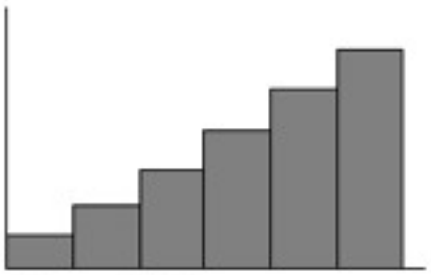
Uniform & symmetrical



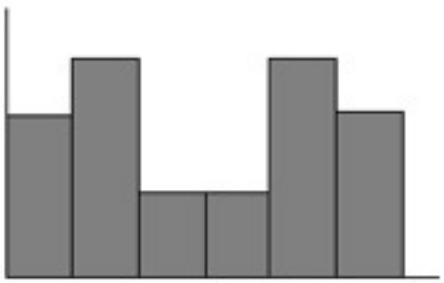
Skewed right



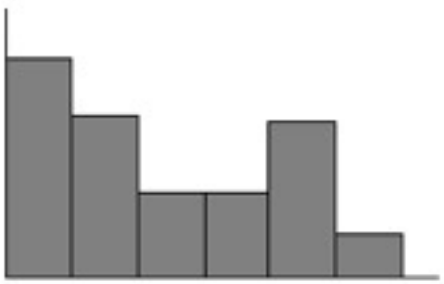
Symmetrical



Skewed left



Bimodal & symmetrical



Bimodal & skewed right

산포도(Scatter Diagram)

- 산포도(scatter diagram)는 두 개의 양적 변인간의 관계를 그래픽으로 나타낸 것.
- 하나의 변인은 수평축에 나타내고, 또 다른 변인은 수직축에 나타냄.
- 점들이 분포된 형태는 변인들의 전반적인 관계를 나타냄.

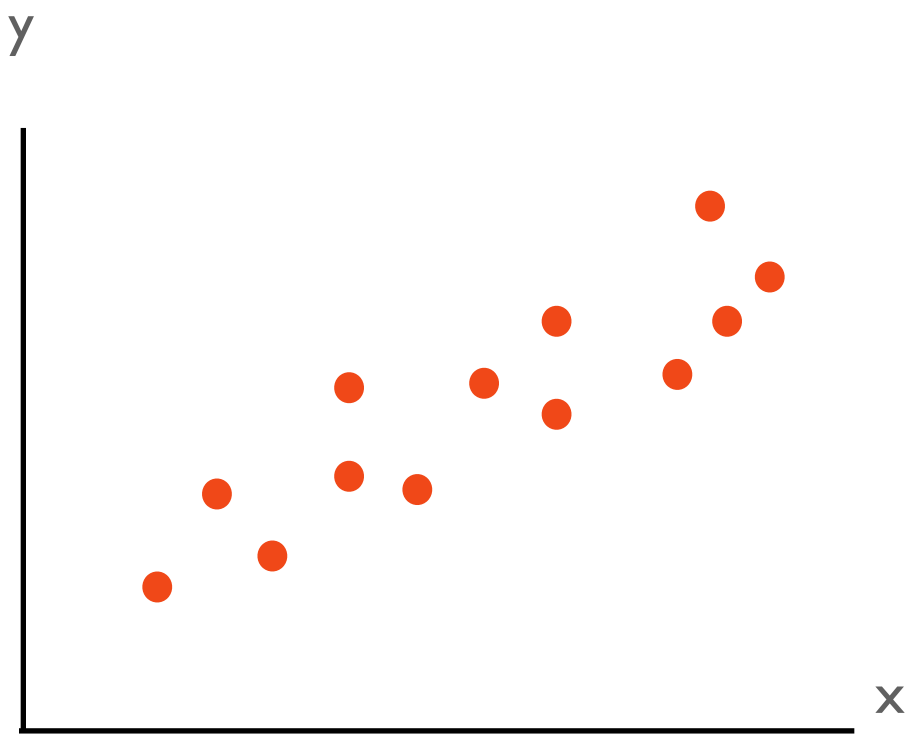
산포도(Scatter Diagram)

정적 관계 (A Positive Relationship)

산포도(Scatter Diagram)

정적 관계 (A Positive Relationship)

산포도(Scatter Diagram)

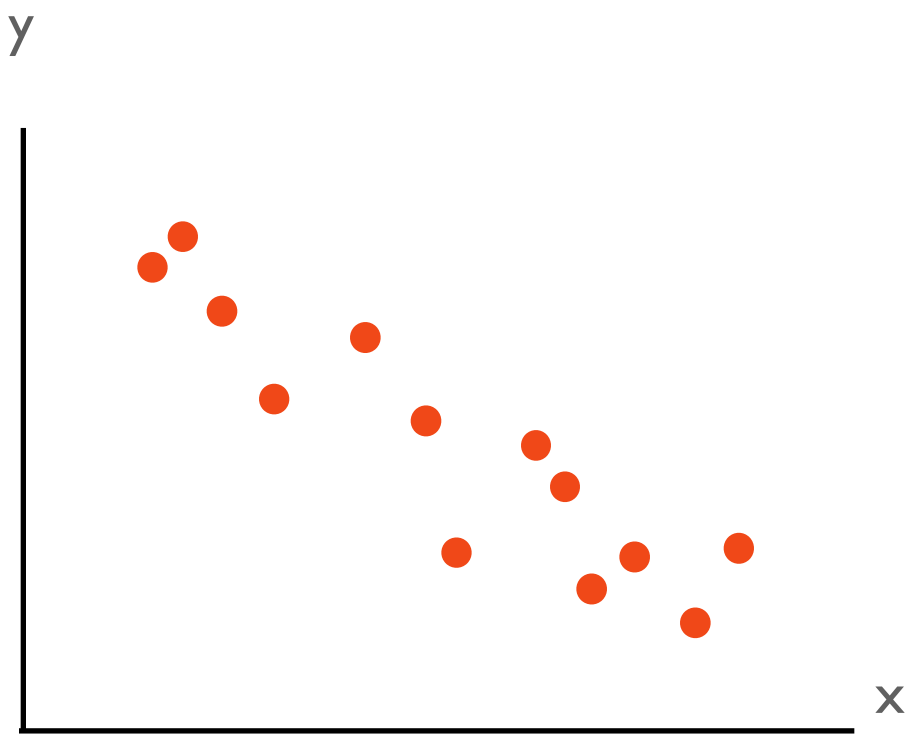


정적 관계 (A Positive Relationship)

산포도(Scatter Diagram)

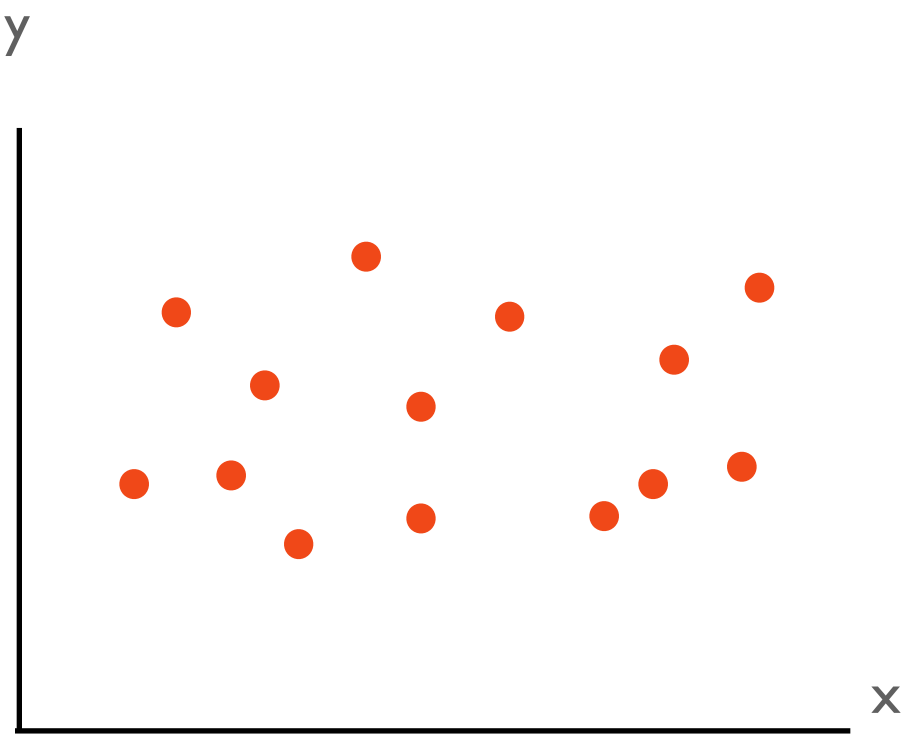
부적 관계(A Negative Relationship)

산포도(Scatter Diagram)



부적 관계(A Negative Relationship)

산포도(Scatter Diagram)



뚜렷한 관계가 없음(No Apparent Relationship)

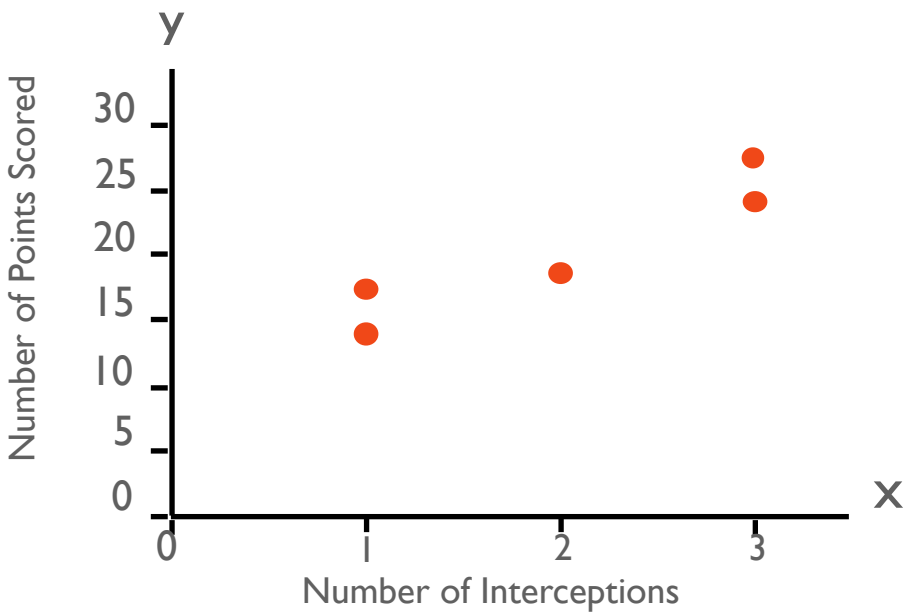
Example: 한양대 축구 팀

- 한양대 미식축구팀의 인터셉트와 득점의 관계.

x = 인터셉트의 수	y = 득점
1	14
3	24
2	18
1	17
3	27

Example: 한양대 축구 팀

- 산포도(Scatter Diagram)



Example: 한양대 축구 팀

- 산포도는 인터셉트와 득점의 정적 관계를 나타냄.
- 높은 득점은 높은 숫자의 인터셉트와 관련이 있었음.
- 모든 점이 직선으로 연결되지 않았으므로 관계가 완벽하다고는 말할 수 없음.

줄기와 잎 그림 (Stem & Leaf Plots)

비교적 적은 양의 자료들은 줄기와 잎(stem-and-leaf) 그림을 이용해서 빠르고 적절하게 설명할 수 있음.

30, 26, 26, 36, 48, 50, 16, 31, 22, 27, 23, 35, 52, 28, 37

Stem (Tens)	Leaf (Units)
1	6
2	66738
3	06157
4	8
5	02

중심경향값
(Measures of Central Tendency)

평균(Mean), $\bar{Y}$	기대값들의 평균 (Weighted center)
중앙값(Median), Md	배열된 값들의 중앙점 (Middle)
최빈값(Mode), Mo	가장 도수가 높은 값 (Most)

평균(Mean)

- 분포의 가중 중심점의 수치.
- 자료 안의 값들의 총합(the sum of the values)을 값들의 개수(the number of values)로 나누어 계산한다.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

= 모집단의 평균 추정치 μ

- Outliers에 의해 영향을 받음.

Example: 자동차 마일리지

5개 차량의 마일리지 평균을 구하라.

30.8, 31.7, 30.1, 31.6, 32.1

Example: 자동차 마일리지

5개 차량의 마일리지 평균을 구하라.
30.8, 31.7, 30.1, 31.6, 32.1

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

Example: 자동차 마일리지

5개 차량의 마일리지 평균을 구하라.
30.8, 31.7, 30.1, 31.6, 32.1

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$= \frac{30.8 + 31.7 + 30.1 + 31.6 + 32.1}{5} = \frac{156.5}{5} = 31.26$$

중앙값(median)

- 나열된 자료의 분포를 동등하게 나눈 중심값
- 측정 값(value)의 50%가 위(혹은 아래)에 존재하는 값.
- N(사례수)이 홀수일 경우 관찰된 중간값이며,
- N(사례수)이 짝수일 경우 두 개의 중앙값을 더한 후 2로 나눈 값으로 나타냄.

Example : Housing Prices

- n 이 홀수인 경우 (7)

{xi} = {house prices}
= {$144,000; 98,000; 204,000; 177,000; 155,000; 316,000; 100,000}

Example : Housing Prices

{xi} = {house prices}

Example : Housing Prices

{xi} = {house prices}

순서대로 배열:

1. $ 98,000
2. 100,000
3. 144,000
4. 155,000
5. 177,000
6. 204,000
7. 316,000

Example : Housing Prices

{xi} = {house prices}

순서대로 배열:

1. $ 98,000
2. 100,000
3. 144,000
4. 155,000
5. 177,000
6. 204,000
7. 316,000

Middle Value
Median = 155,000

Example : Housing Prices

n 이 짝수인 경우 (10)

{xi } = {house prices}
= {$144,000; 98,000; 204,000; 177,000; 155,000; 316,000;
100,000; 177,000; 177,000; 170,000}

Example : Housing Prices

순서대로 배열:

1. $ 98,000
2. 100,000
3. 144,000
4. 155,000
5. 170,000
6. 177,000
7. 177,000
8. 177,000
9. 204,000
10. 316,000

Example : Housing Prices

순서대로 배열:

1. $ 98,000
2. 100,000
3. 144,000
4. 155,000
5. 170,000
6. 177,000
7. 177,000
8. 177,000
9. 204,000
10. 316,000

Middle Values

Example : Housing Prices

순서대로 배열:

1. $ 98,000
2. 100,000
3. 144,000
4. 155,000
5. 170,000
6. 177,000
7. 177,000
8. 177,000
9. 204,000
10. 316,000

Middle Values

Median = {170,000 + 177,000}/2
= 173,500

Quiz: 중앙값을 찾아라.

Quiz: 중앙값을 찾아라.

13명 의사의 월급 (x$1000)

Quiz: 중앙값을 찾아라.

13명 의사의 월급 (x$1000)

127 132 138 141 144 146 152 154 165 171 177 192 241

Quiz: 중앙값을 찾아라.

13명 의사의 월급 (x$1000)

127 132 138 141 144 146 152 154 165 171 177 192 241

n = 13,

Quiz: 중앙값을 찾아라.

13명 의사의 월급 (x$1000)

127 132 138 141 144 146 152 154 165 171 177 192 241

n = 13,
중앙값(median)은 7번째 값

Quiz: 중앙값을 찾아라.

13명 의사의 월급 (x$1000)

127 132 138 141 144 146 152 154 165 171 177 192 241

n = 13,
중앙값(median)은 7번째 값

Md=152

최빈값(The Mode)

분포의 최빈값(Mode, Mo)은 가장 빈번하게 발생하는 사례값을 나타냄.

Example: Sample Mode

n = 65

10 0
11 0
12 00
13 000
14 0000
15 0000000
16 000000000
17 00000000
18 000000
19 00000
20 000
21 000
22 000
23 00
24 000
25 00
26 0
27 0
28
29 0

Example: Sample Mode

n = 65

값16은 9번 나타남.
따라서

Mo = 16

10 0
11 0
12 00
13 000
14 0000
15 0000000
16 000000000
17 00000000
18 000000
19 00000
20 000
21 000
22 000
23 00
24 000
25 00
26 0
27 0
28
29 0

분산도
(Measures of Variation)

- 범위(Range) : 최고값과 최저값의 차이
- 변량(Variance) : 편차(평균에서 떨어진 정도)를 제곱한 값들의 평균값
- 표준편차(Standard Deviation) : 변량의 제곱근

범위(The Range)

범위 = 최대값 - 최저값

범위(The Range)

범위 = 최대값 - 최저값

Example : 의사들의 월급 (in thousands of dollars)

127 132 138 141 144 146 152 154 165 171 177 192 241

범위(The Range)

범위 = 최대값 - 최저값

Example : 의사들의 월급 (in thousands of dollars)

127 132 138 141 144 146 152 154 165 171 177 192 241

범위 = 241 - 127 = $114k ($114,000)

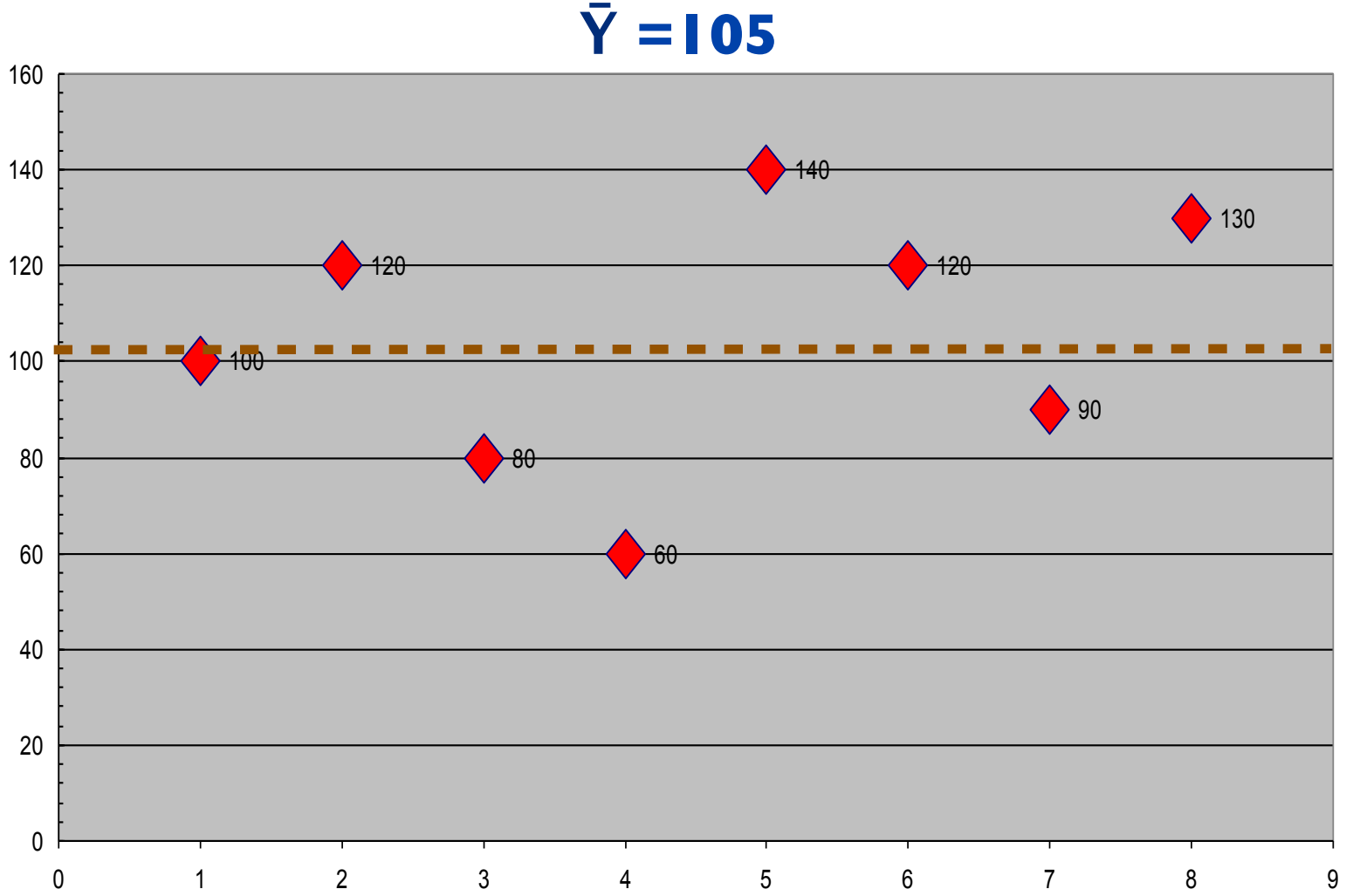
변량과 표준편차 (Variance & Standard

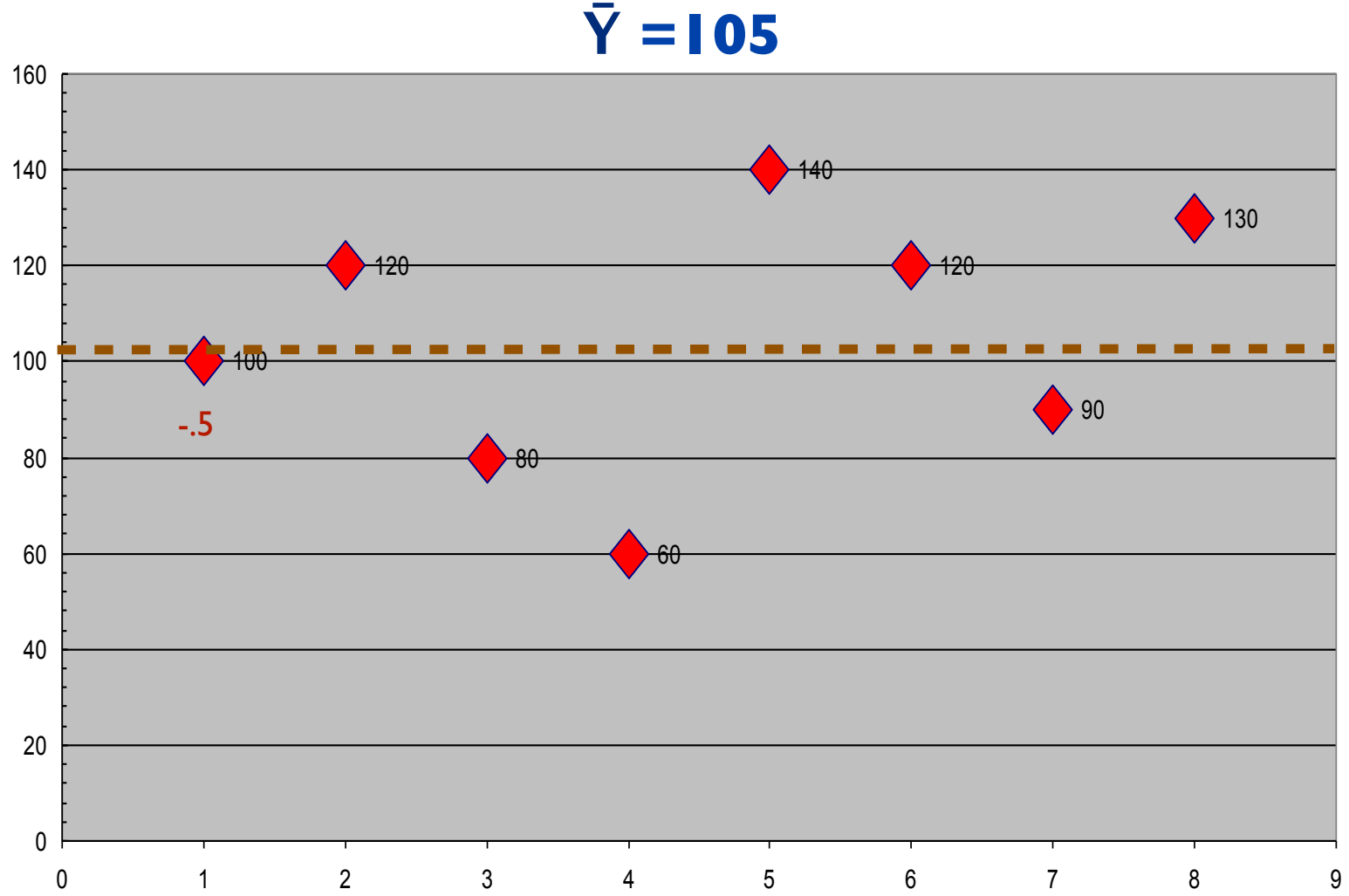
- 변량(Variance): 편차제곱의 평균

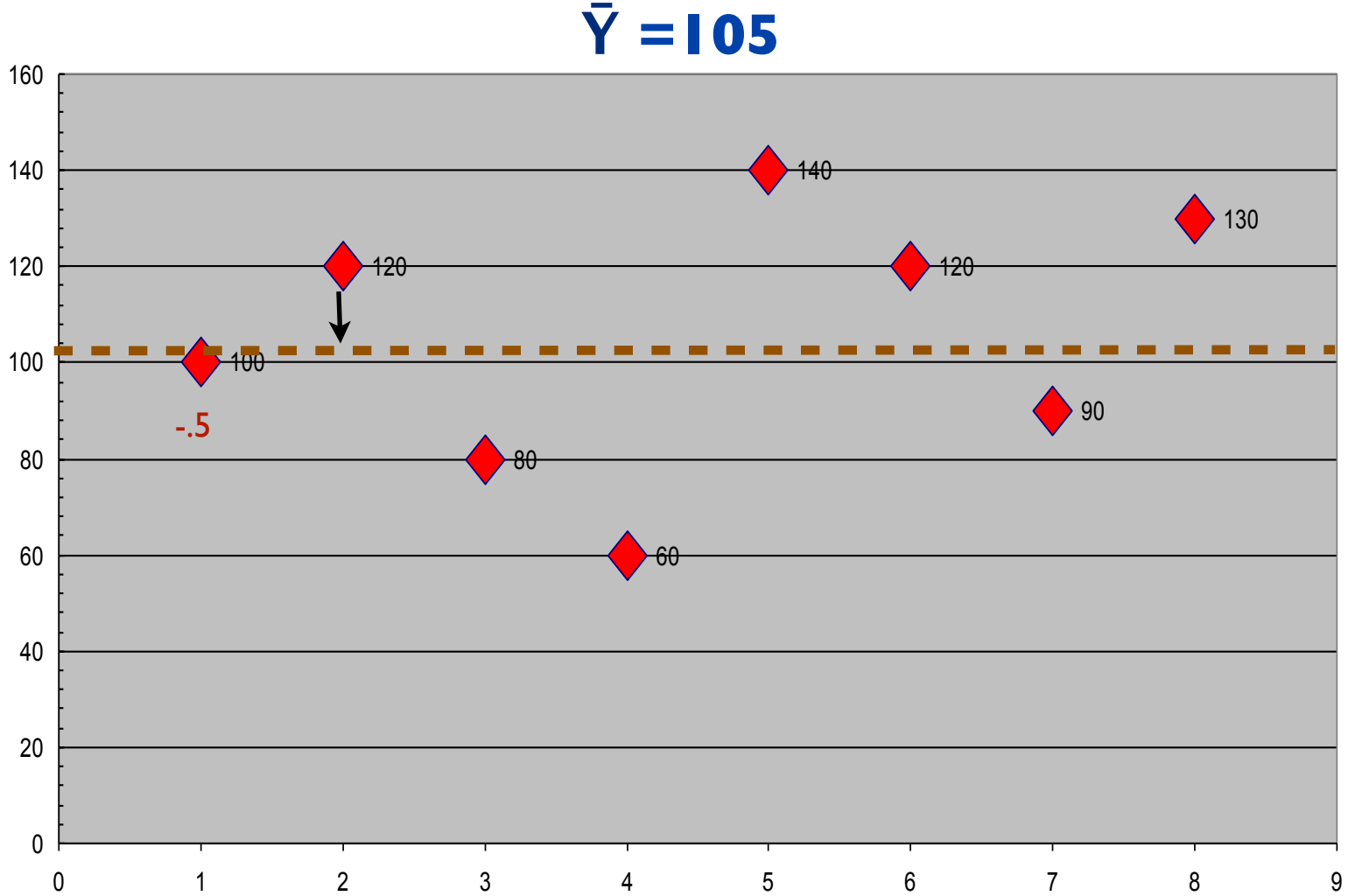
  $s^2 = \Sigma\{Yi - Y\}^2 / n - 1 =$ 변량 (Variance)

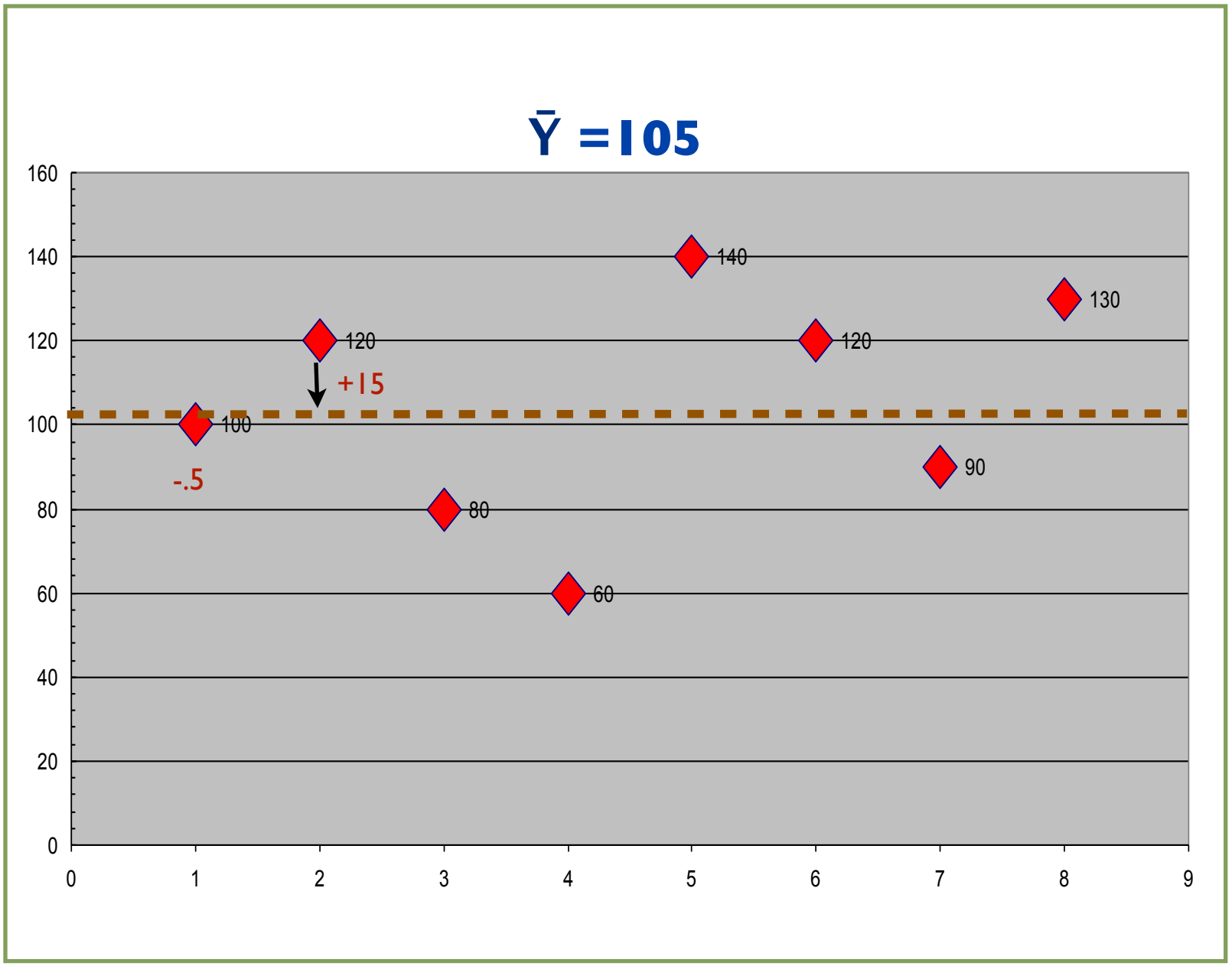
- 표준편차(Standard Deviation) : 변량의 양의 제곱근
  - 변량의 밀도(density of dispersion)에 대한 정량적 측정값
  - 극단값(extreme values)에 의해 영향을 적게 받음.

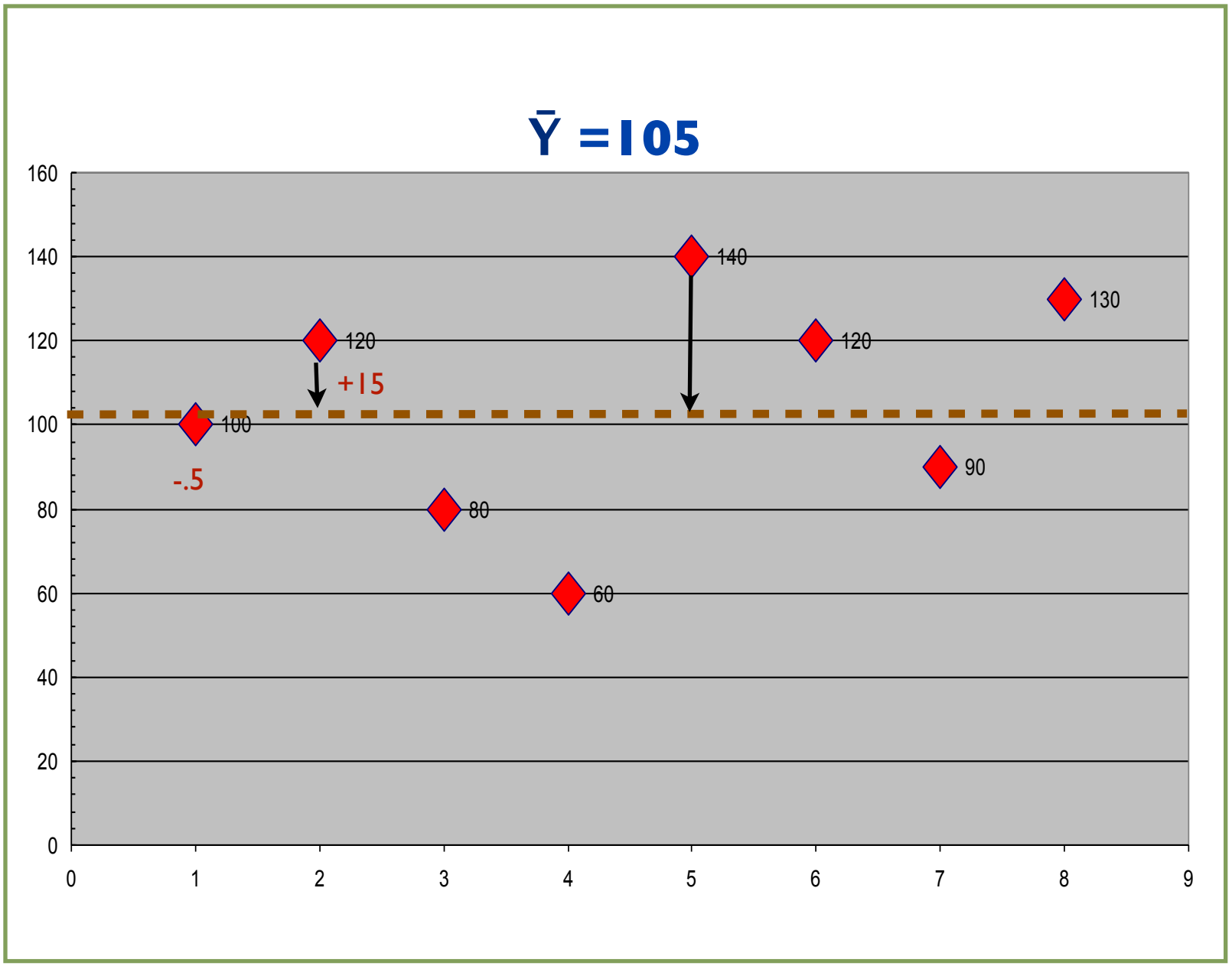
  $s = \sqrt{s^2} =$ 표준편차 (Standard Deviation)

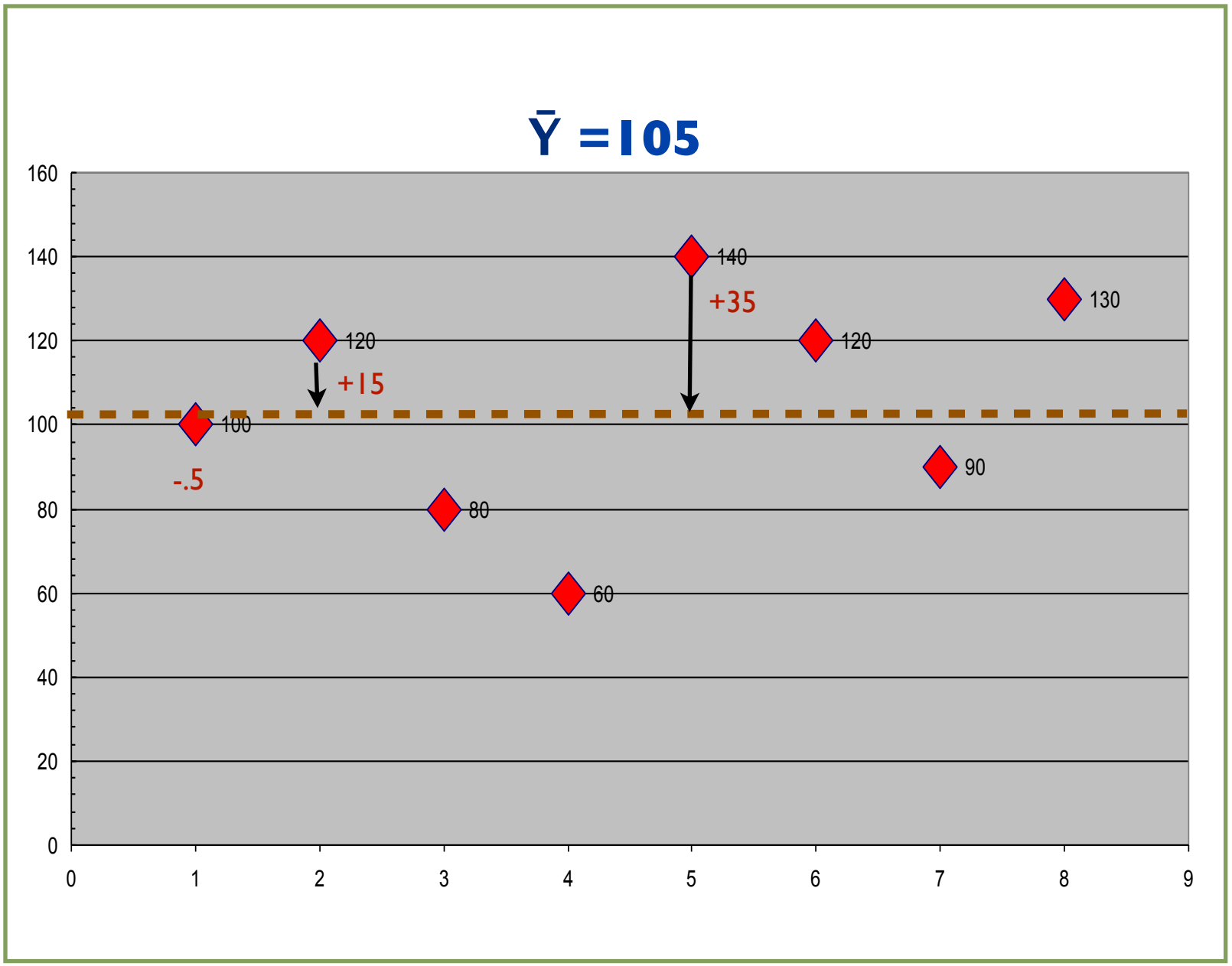

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
140
130
120
120
100
90
80
60

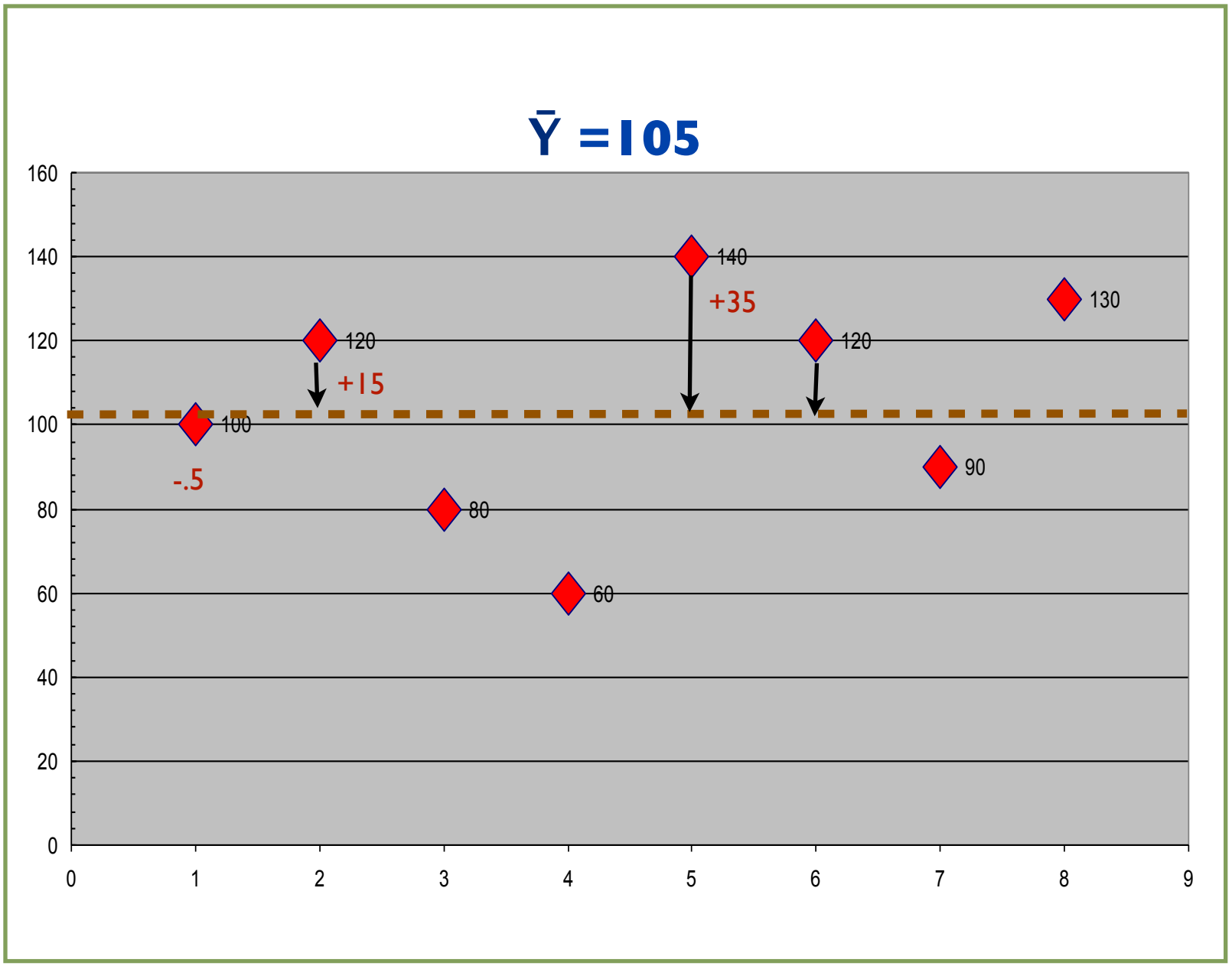

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
140
130
120
120
100
90
80
60
-.5

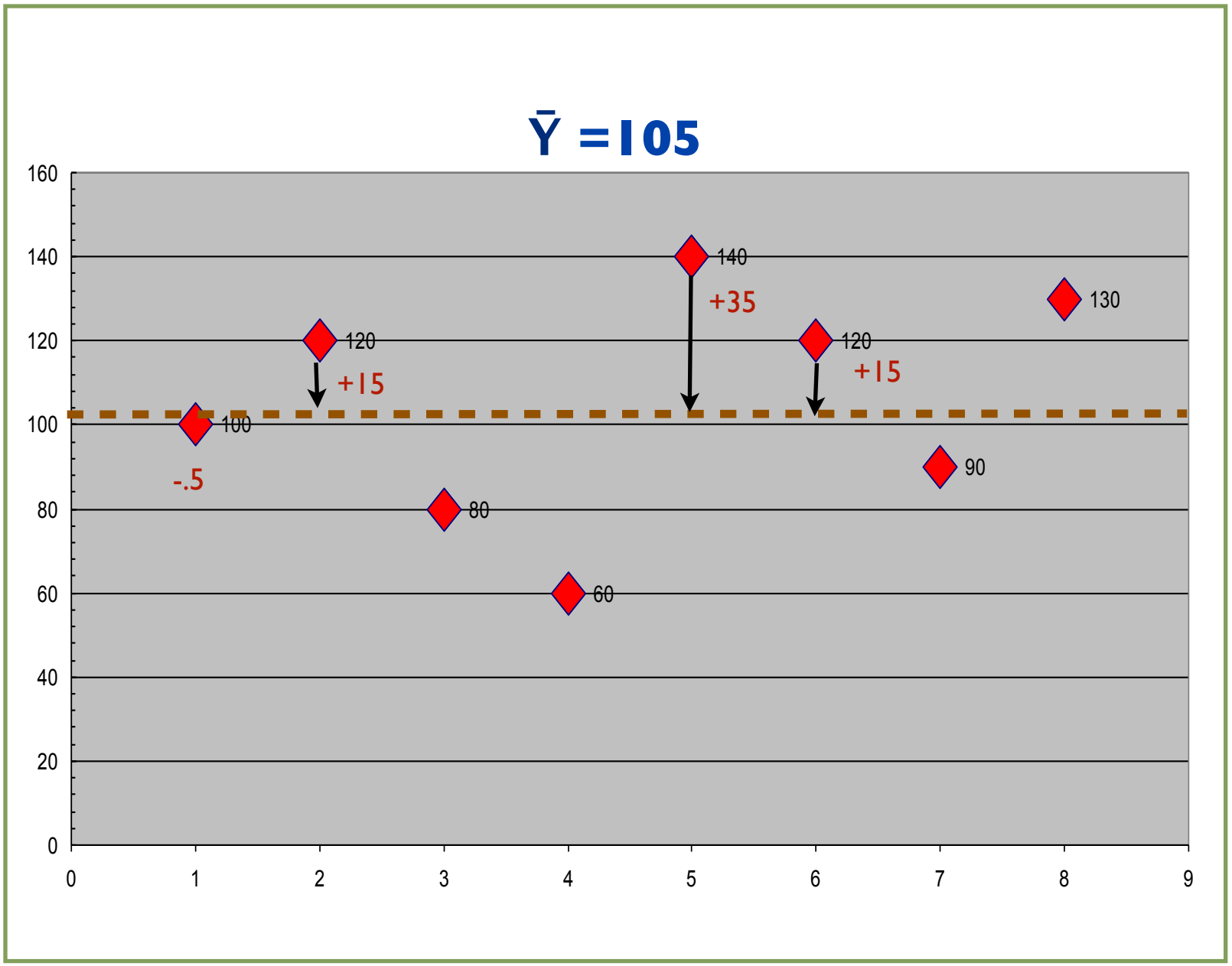

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
140
130
120
120
100
90
80
60
-.5

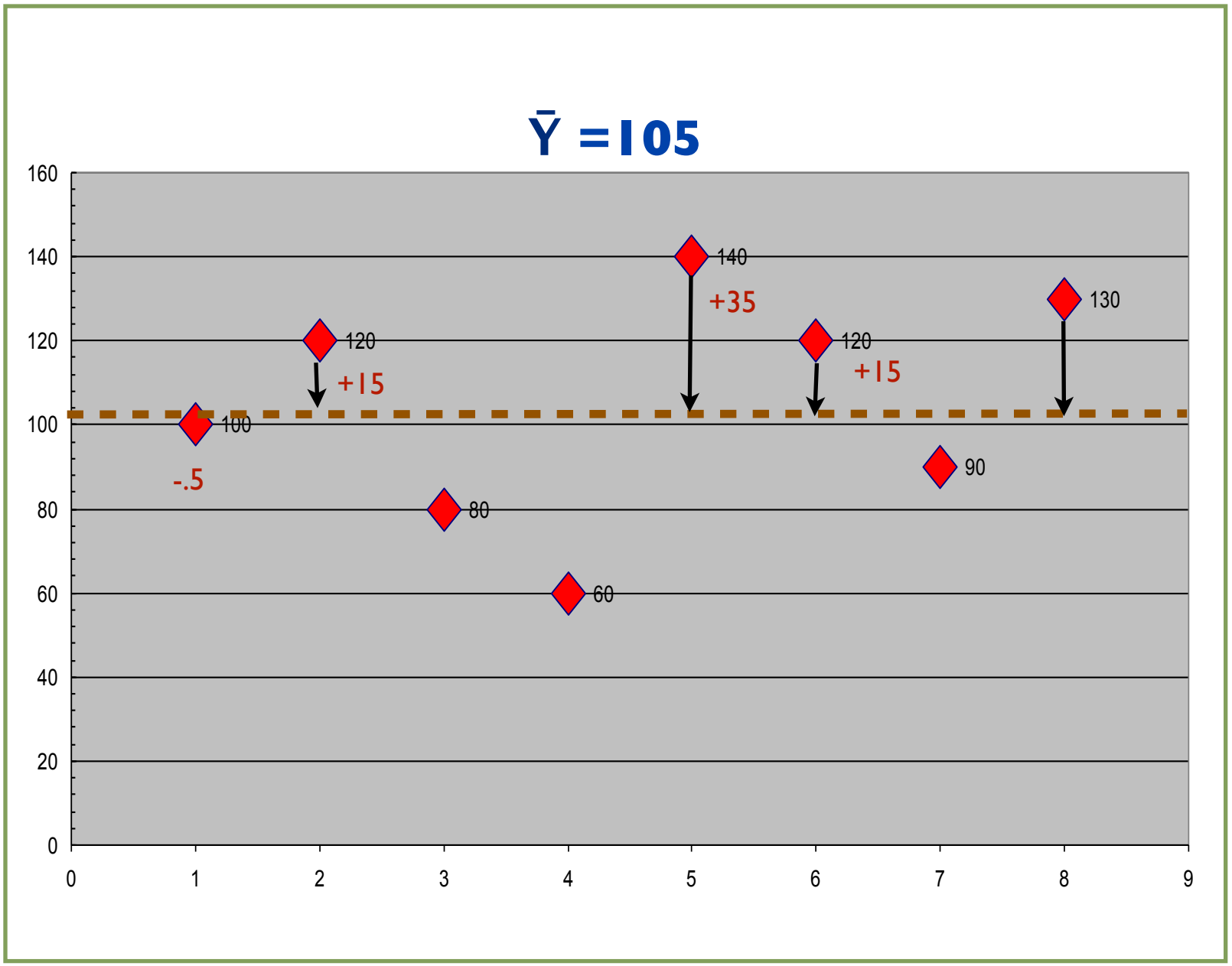

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
140
130
120
120
100
90
80
60
+15
-.5

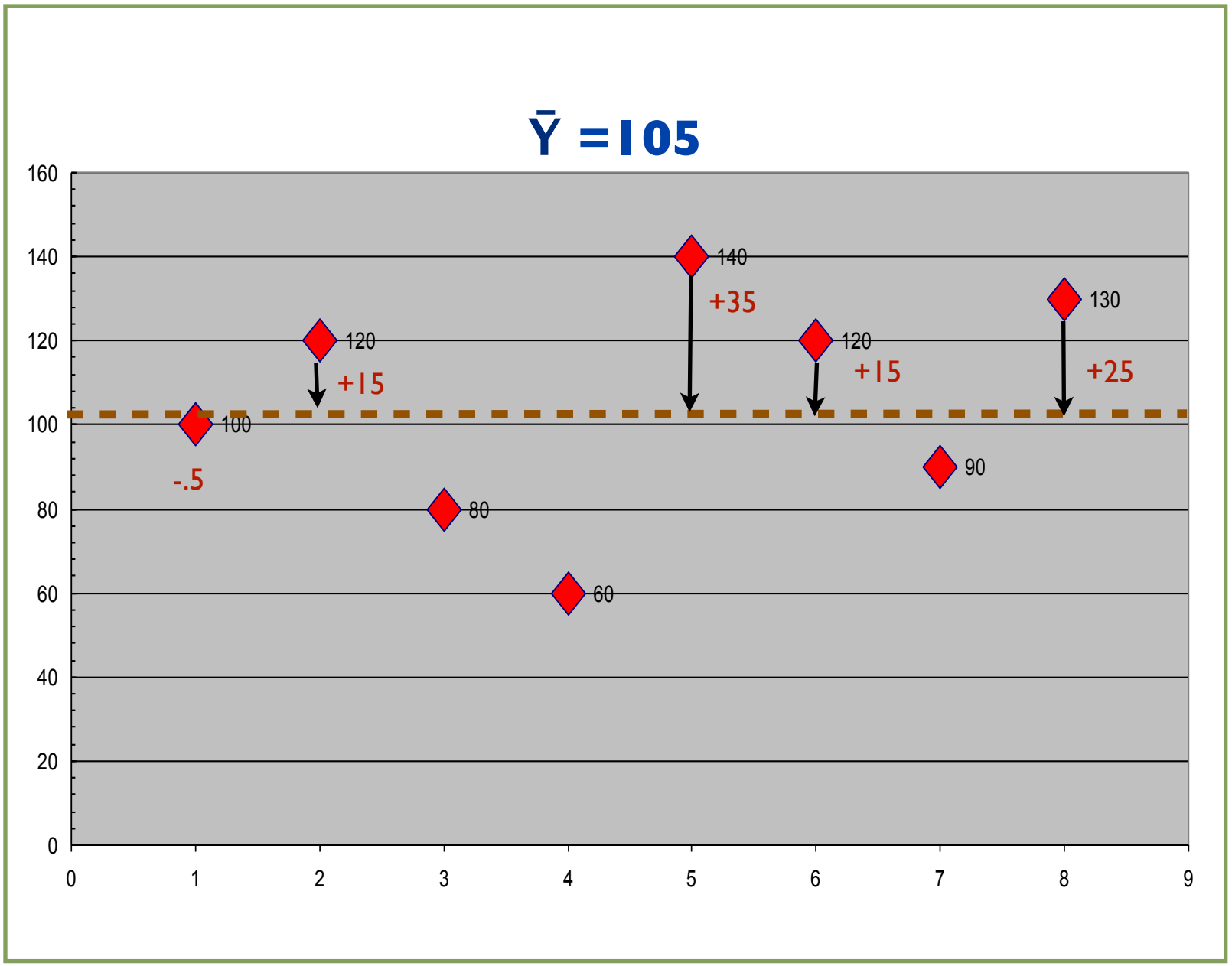

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
120
140
120
130
90
80
60
+15
-.5

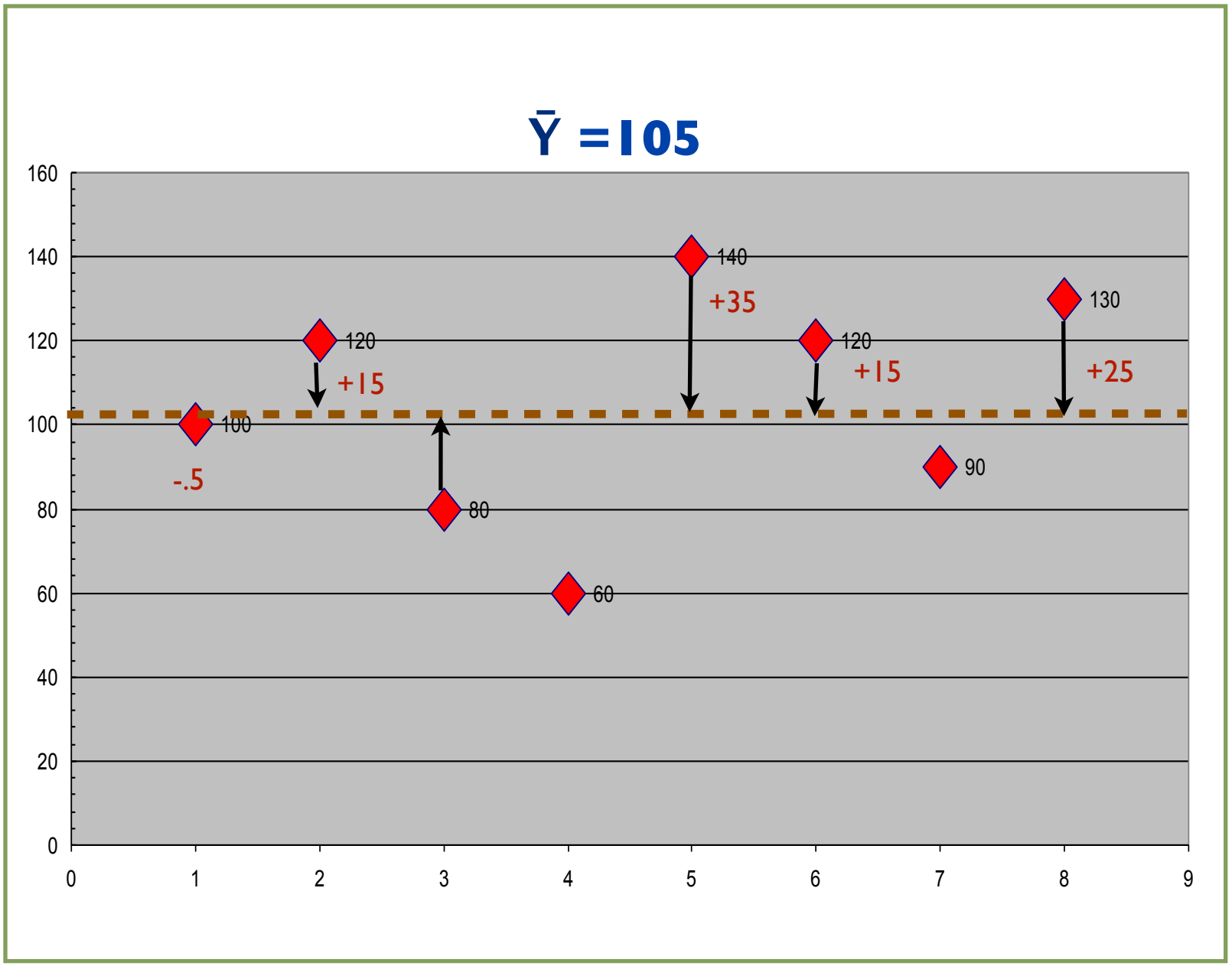

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
140
+35
120
120
130
+15
100
-.5
90
80
60

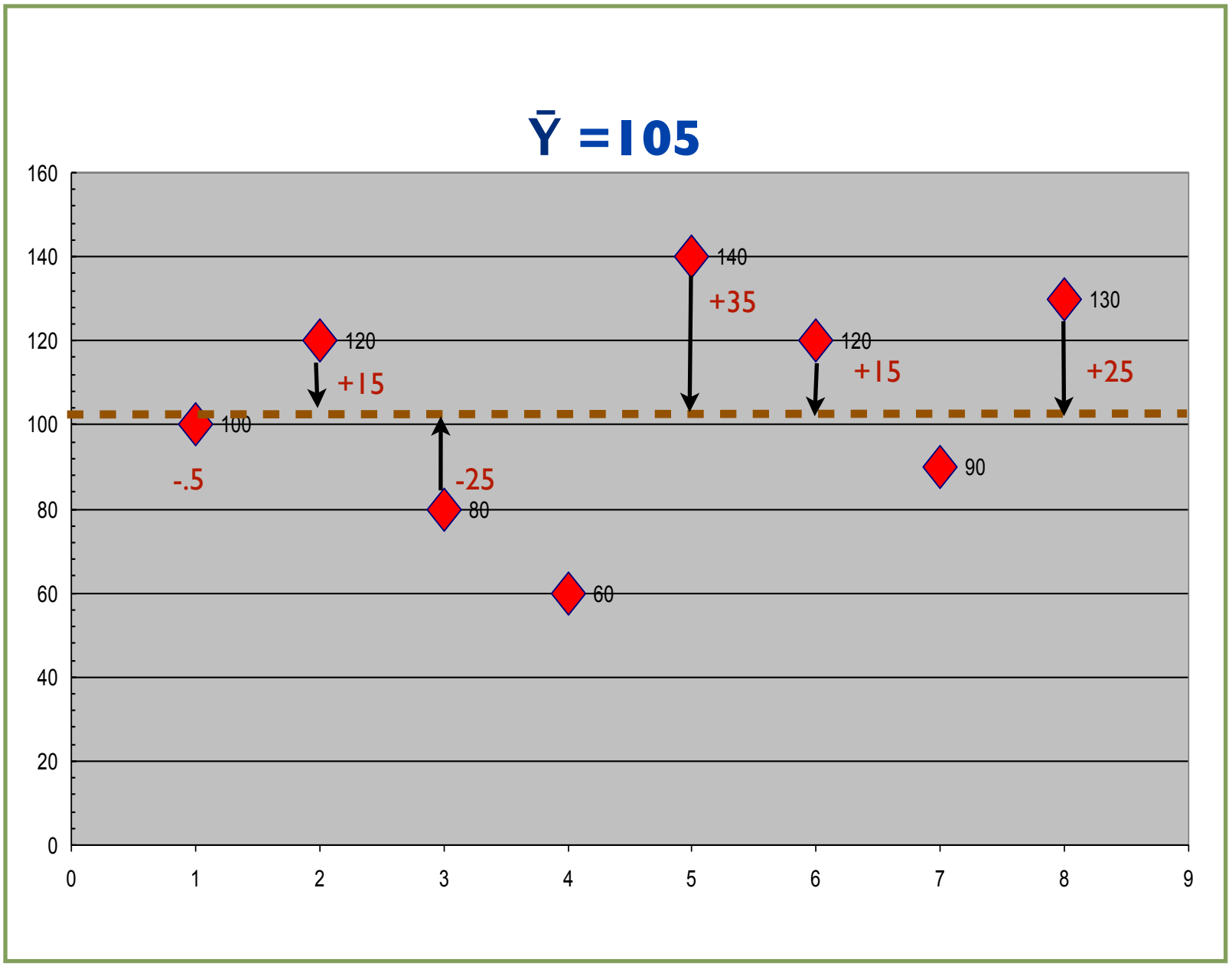

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
120
140
80
60
120
130
90
-.5
+15
+35

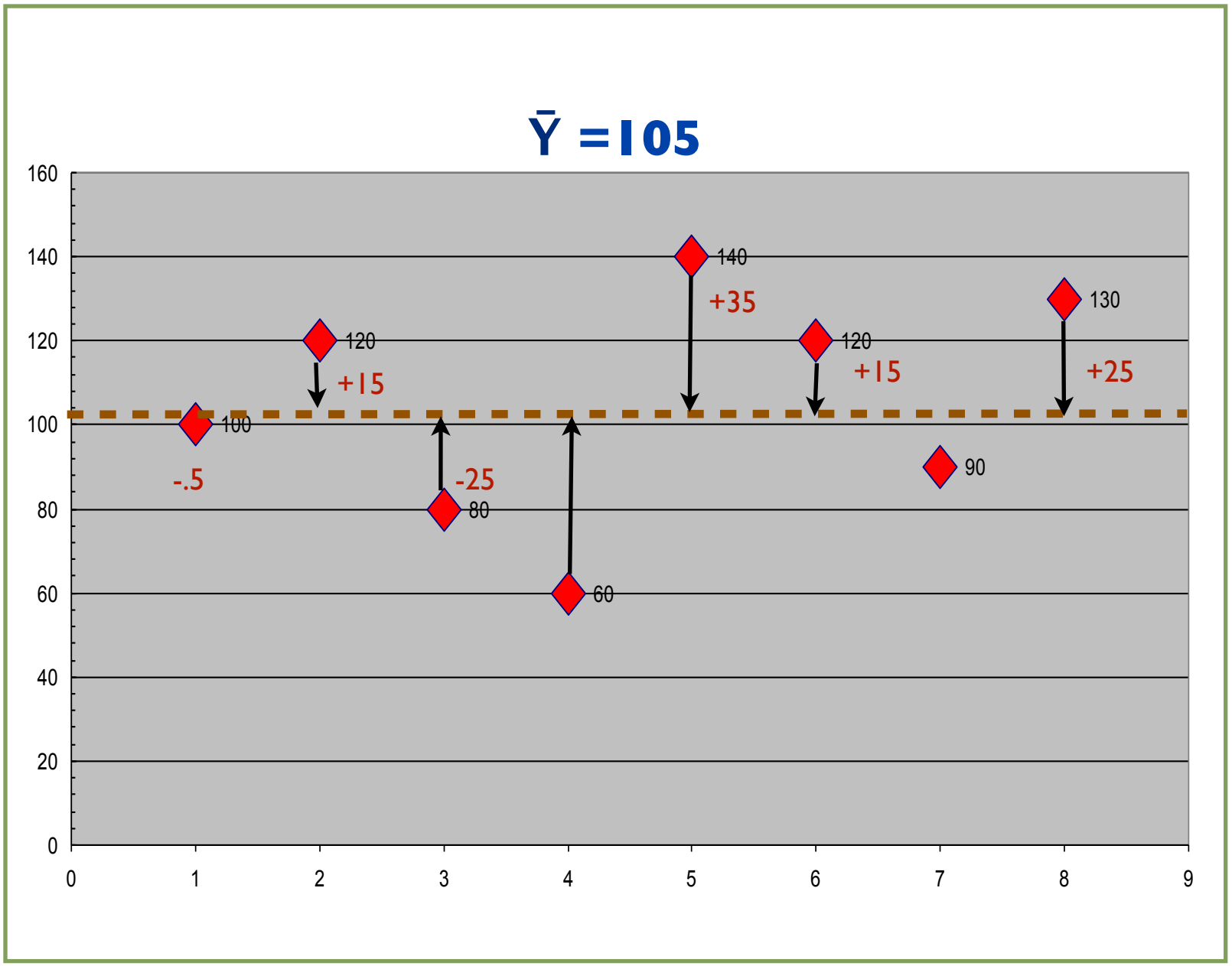

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
120
140
120
130
90
80
60
-.5
+15
+35
+15

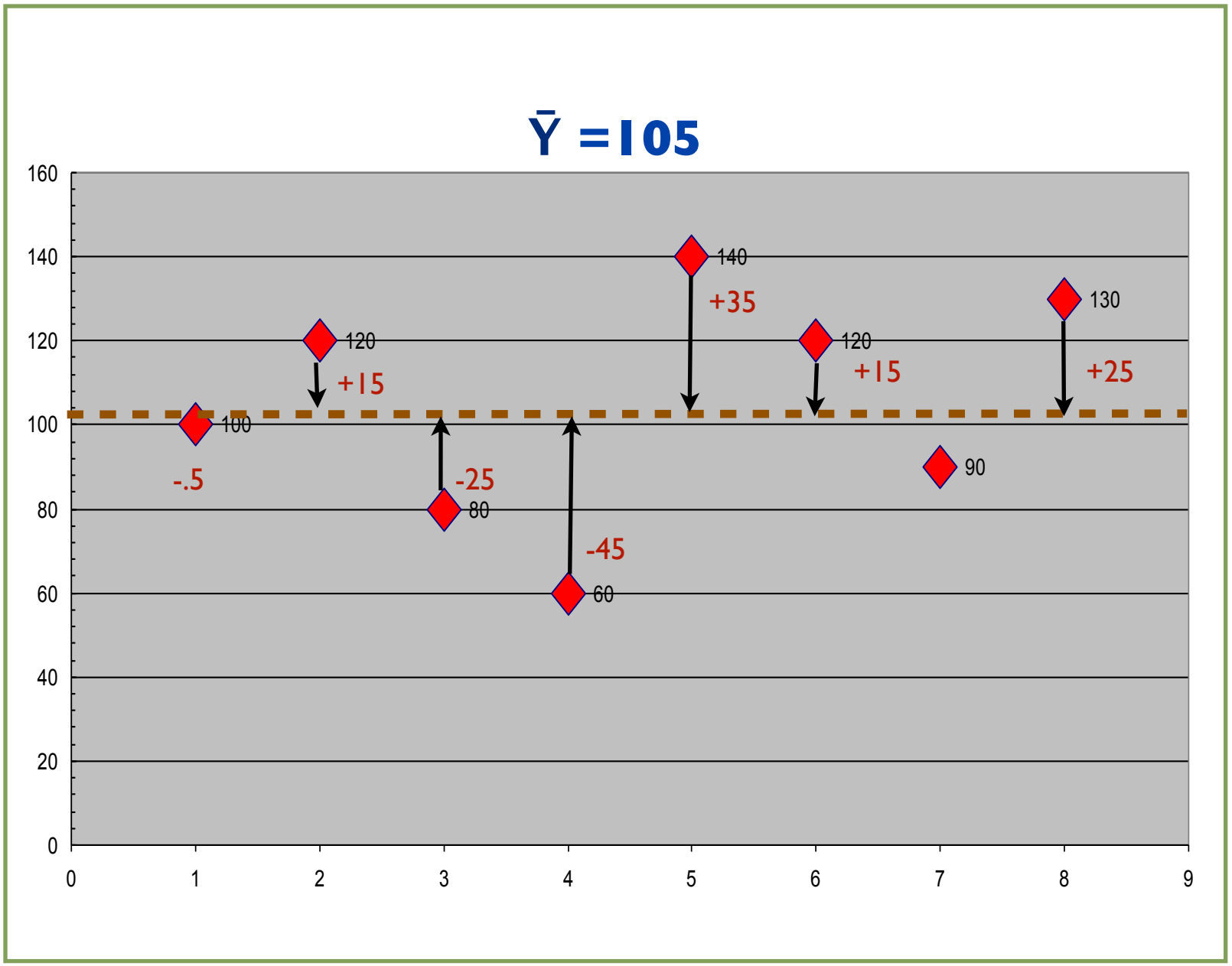

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
120
+15
140
+35
120
+15
130
90
80
60
-.5

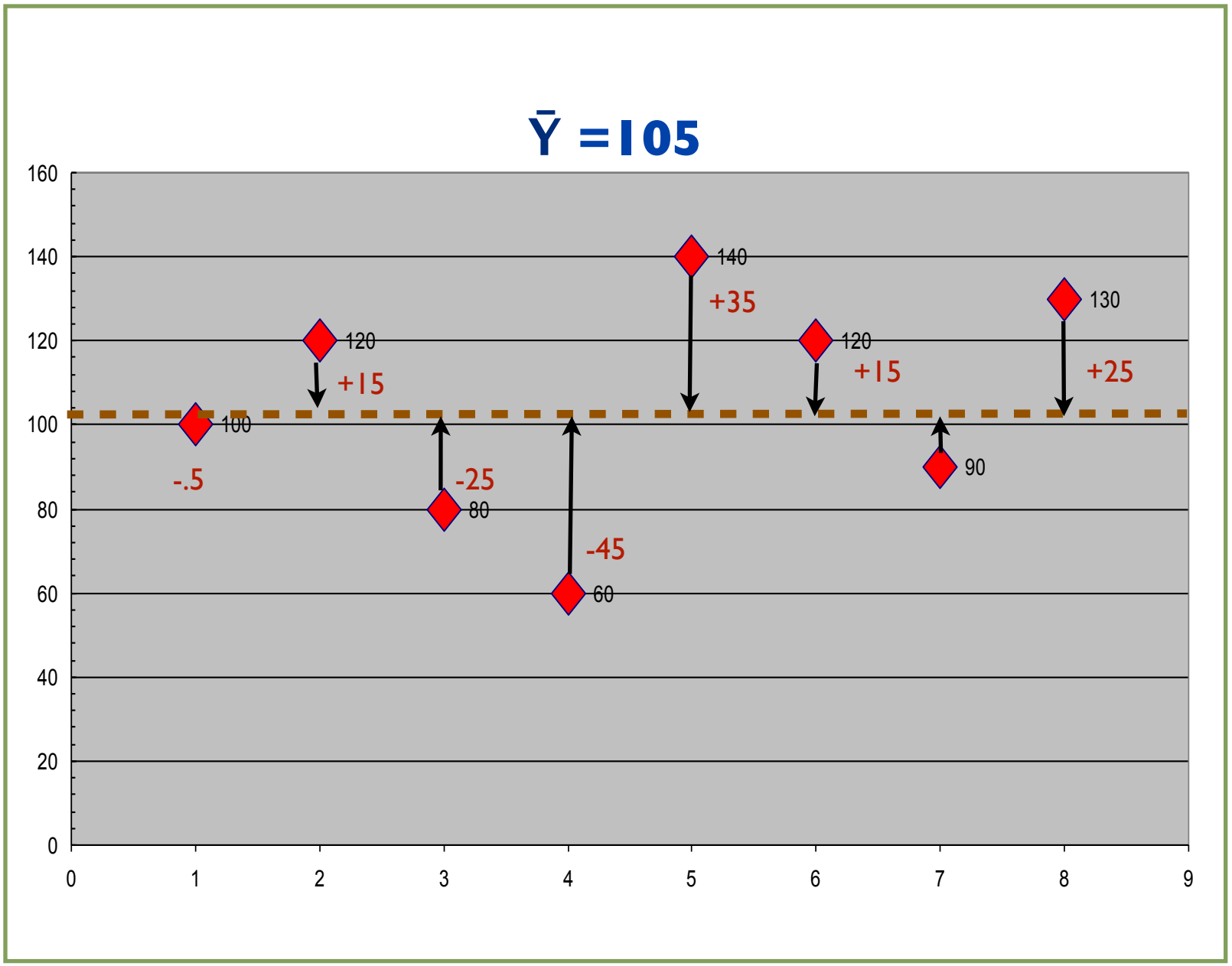

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
120
140
80
60
90
120
130
-.5
+15
+35
+15
+25

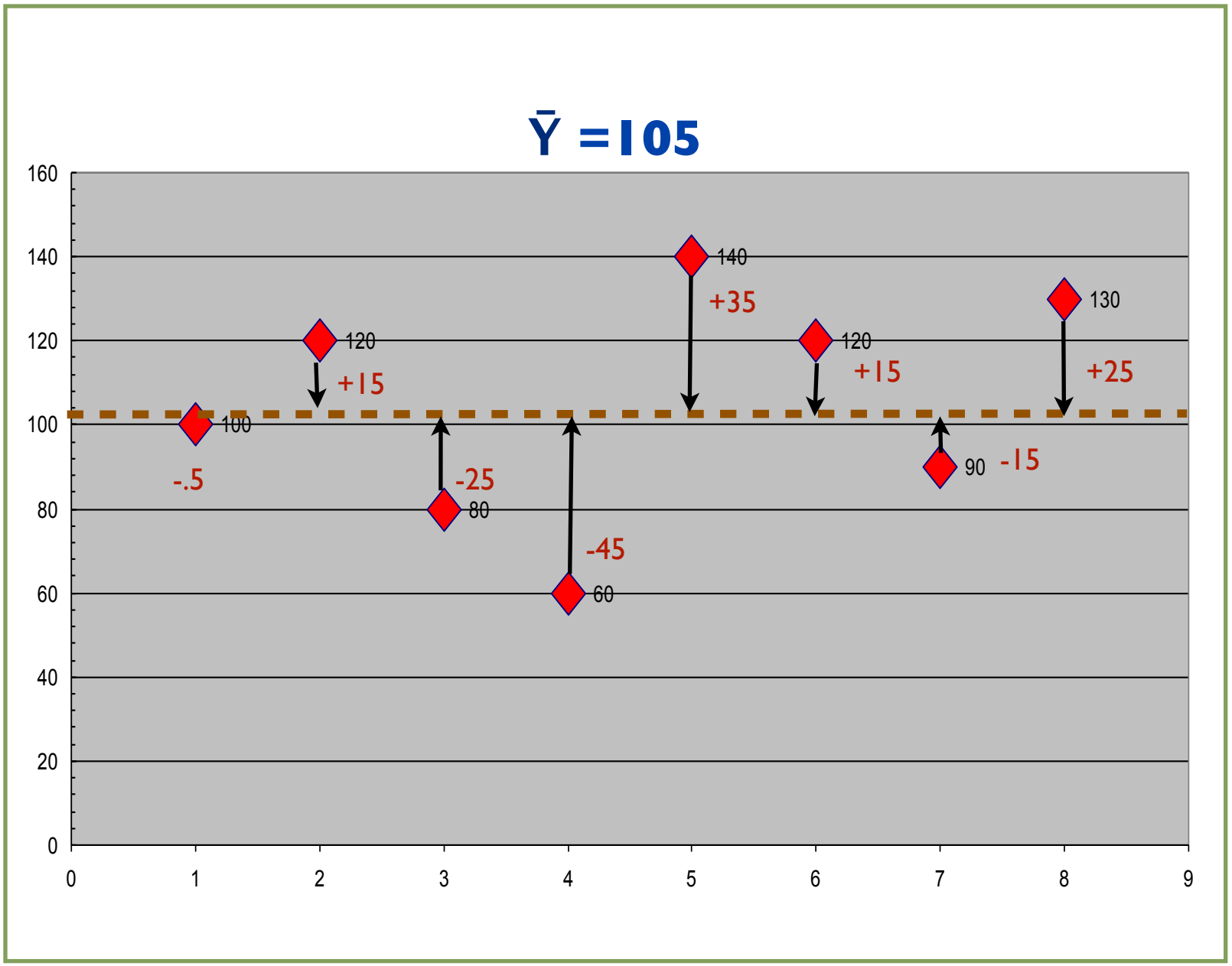

$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
-.5
120
+15
80
140
+35
60
120
+15
90
130
+25


$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
120
+15
-.5
-25
80
60
140
+35
120
+15
90
130
+25


$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
120
+15
-.5
-25
80
60
140
+35
120
+15
90
130
+25


$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
-.5
120
+15
80
-25
60
-45
140
+35
120
+15
90
130
+25


$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
-.5
120
+15
80
-25
60
-45
140
+35
120
+15
90
130
+25


$\bar{Y}$ =105
160
140
120
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
100
-.5
120
+15
80
-25
60
-45
140
+35
120
+15
90
-15
130
+25

표준편차 s를 구하라.

- $s^2 = \Sigma (Yi - \bar{Y})^2 / n-1 =$
- $s =$

- 윤석이는 홍보조사실습을 듣는 학생들의 주당 음주횟수를 조사하였다. 결과는 다음과 같다:

- 윤석이 데이터의 표준편차(standard deviation)는?

Drinking per week	Frequency
0	4
1	12
2	8
3	2
4	1
5	2
6	1
Total	30

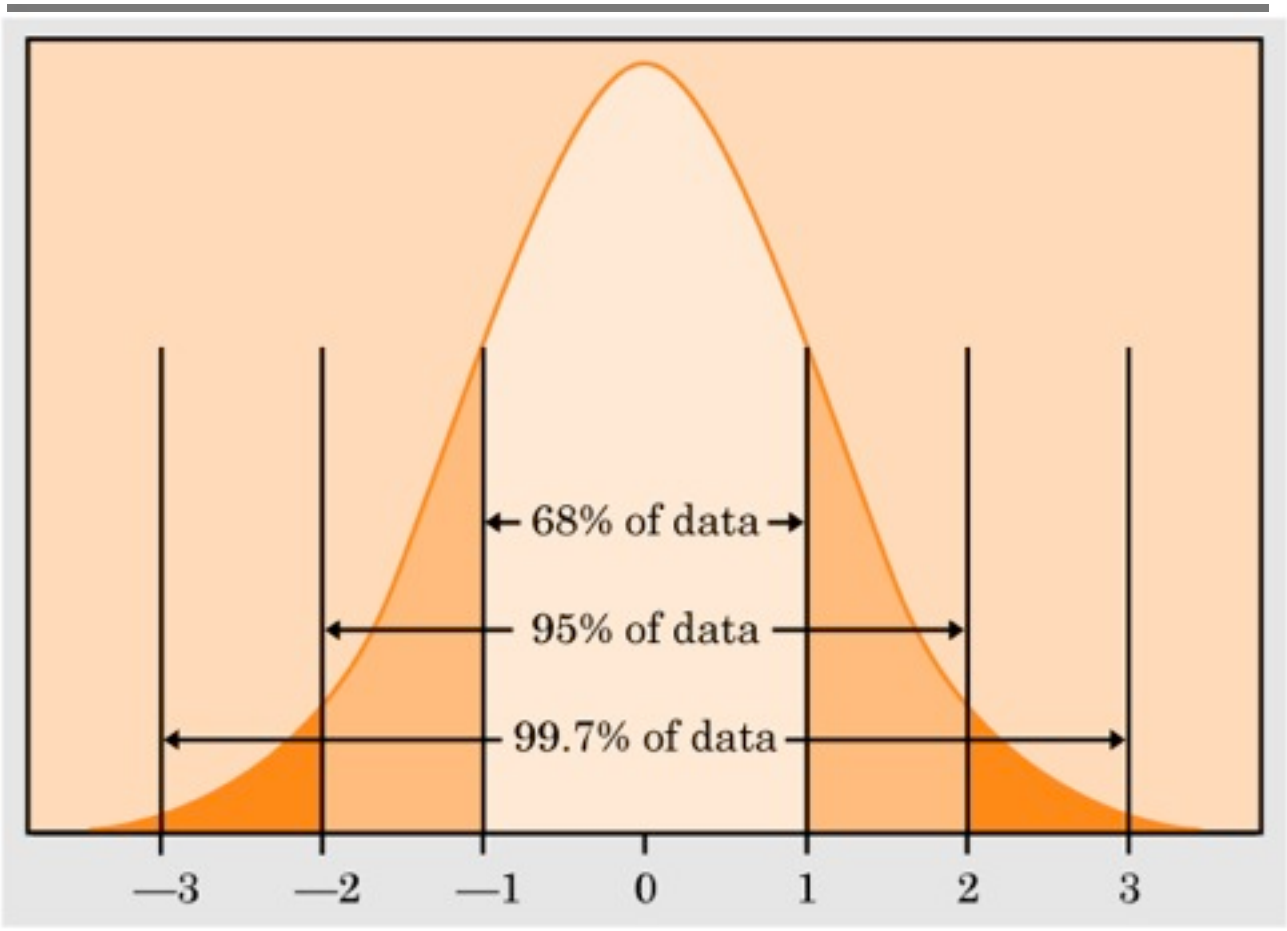
X	f	fX	$X - \bar{X}$	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
0	4	0	-1.8	3.24	12.96
1	12	12	-0.8	0.64	7.68
2	8	16	0.2	0.04	0.32
3	2	6	1.2	1.44	2.88
4	1	4	2.2	4.84	4.84
5	2	10	3.2	10.24	20.48
6	1	6	4.2	17.64	17.64
	$\Sigma f = 30$	$\Sigma fX = 54$			$\Sigma f(X - \bar{X})^2 = 66.80$

$$\bar{X} = \frac{\Sigma fx}{\Sigma f} = \frac{54}{30} = 1.8$$

$$s = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}} = \sqrt{\frac{66.80}{30}} = \sqrt{2.2266\ldots} = 1.49 \text{ correct to 2 decimal places}$$

The Empirical Rule

- 만약 자료의 히스토그램이 대략의 정규분포를 이루고 있다면,
  - 자료의 68% 가 $\bar{Y} - s$ 와 $\bar{Y} + s$ 사이에 존재할 것임.
  - 자료의 95% 가 $\bar{Y} - 2s$ 와 $\bar{Y} + 2s$ 사이에 존재할 것임.
  - 거의 모든 자료(99.7%)가 $\bar{Y} - 3s$ 와 $\bar{Y} + 3s$ 사이에 존재할 것임.



The Empirical Rule

Example: 12세 아동들의 IQ 점수

- 정규분포를 이루고 있는 12세 아동 IQ점수의 평균은 100이고 표준편차는 16이다.
  - 중앙 68%에 해당하는 점수의 구간을 계산하여라. 95%, 99.7%의 구간을 계산하여라.
  - IQ점수가 100보다 높은 아이들은 전체의 몇 %인가? 116, 132보다 높은 아이들은 몇 %인가?
  - IQ점수가 116보다 낮은 아이들은 전체의 몇 %인가? 100, 84보다 낮은 아이들은 몇 %인가?

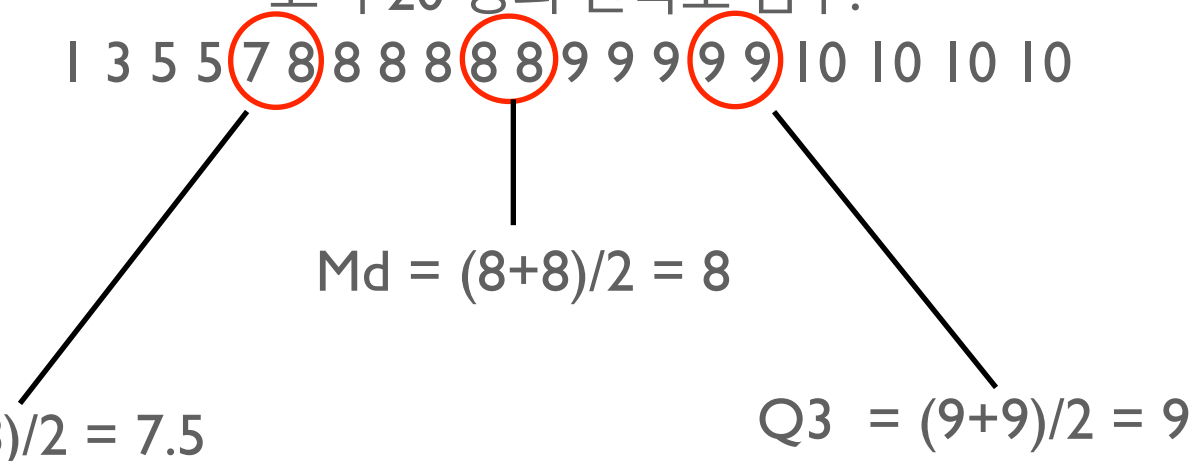
사분위간 범위(Interquartile Range)

- 사분위수(Quartiles)
  - 제1사분위수 Q1 는 제25백분위수를 말함.
  - 제2사분위수 Q2 (중앙값) Md 는 제50백분위수를 말함.
  - 제3사분위수 Q3 는 제75백분위수를 말함.
  - 사분위간 범위(Inter Quartile Range:IQR)는 Q3 - Q1.
- 이상치(Outlier)는 상위 사분위 1.5 IQR 이상이거나 하위 사분위 1.5 IQR이하로 나타난 관찰치.

Example: 사분위수(Quartiles)

고객 20 명의 만족도 점수:

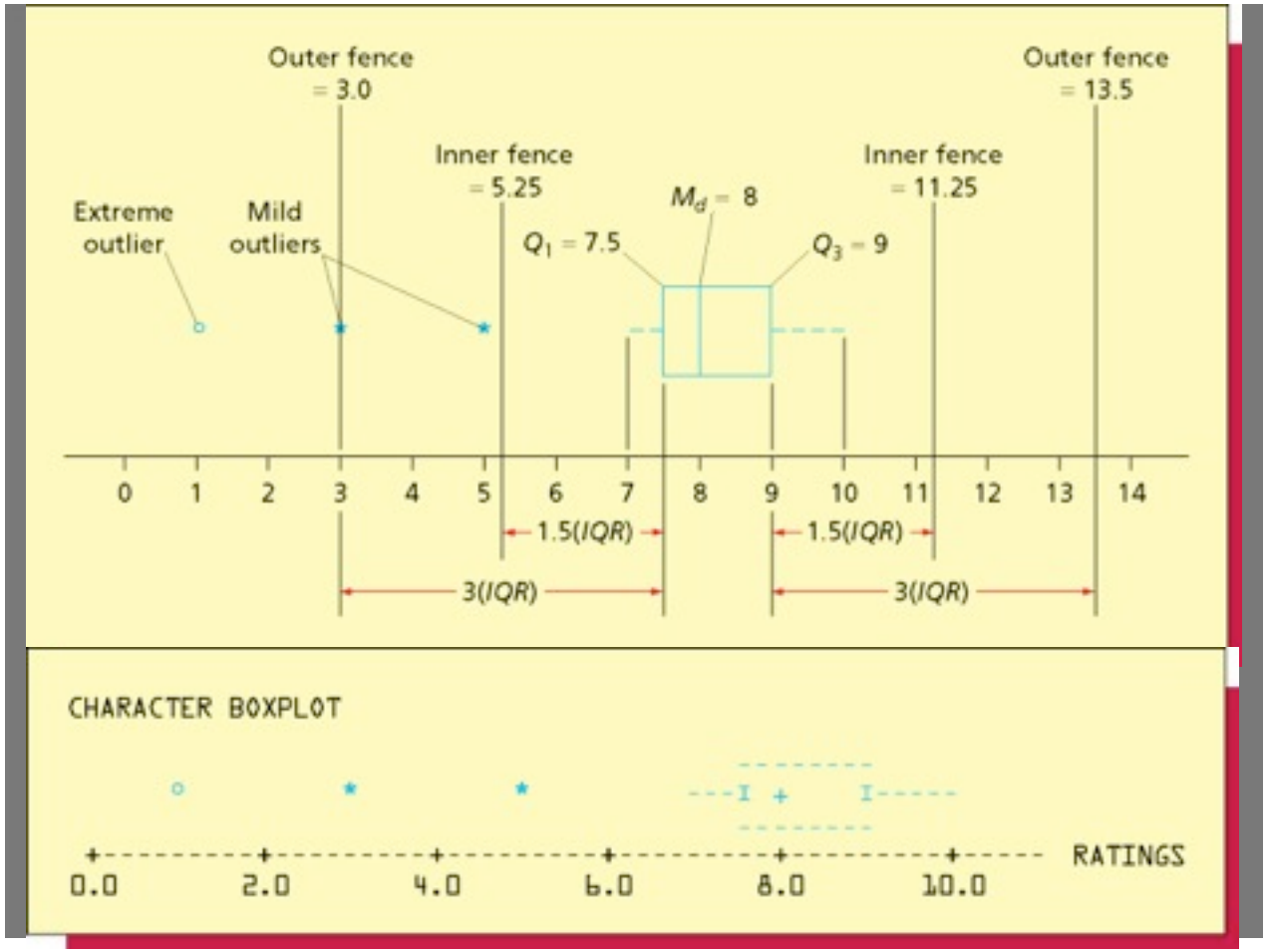
1 3 5 5 7 8 8 8 8 8 8 9 9 9 9 9 10 10 10 10



IQR = Q3 - Q1 = 9 - 7.5 = 1.5

상자 플롯(Box and Whiskers Plots)

- 자료의 중심경향과 분산을 그래픽으로 요약
- 자료의 사분위수(quartiles)와 범위(range), 이상치(outliers)를 그린 것.



상자 플롯
(Box and Whiskers Plots)


Q&A