Lecture Notes for

Introduction to Probability and Statistics

Hanyang University, Ansan, 425-791, Korea

calltimy@hanyang.ac.kr

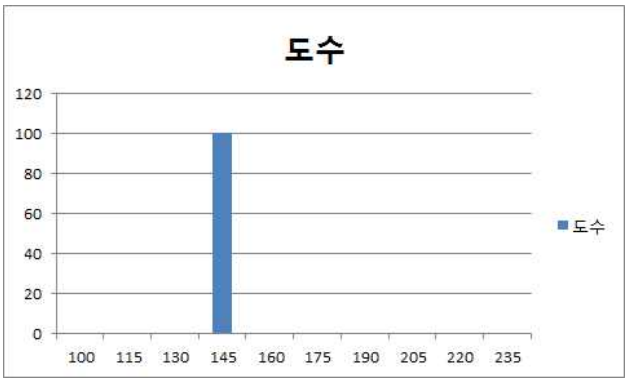
Probability and Statistics / 확률과 통계

강의노트 02

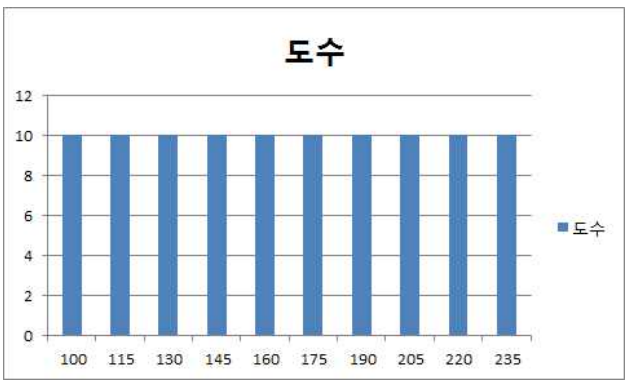
데이터분석 2

14. 산포도 : 데이터가 대표값에서 얼마나 멀리 떨어져 있는지를 나타내는 정도

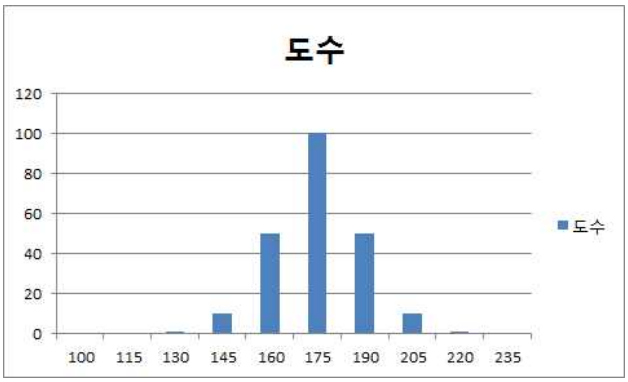
● 모든 학생의 몸무게가 145 파운드라면



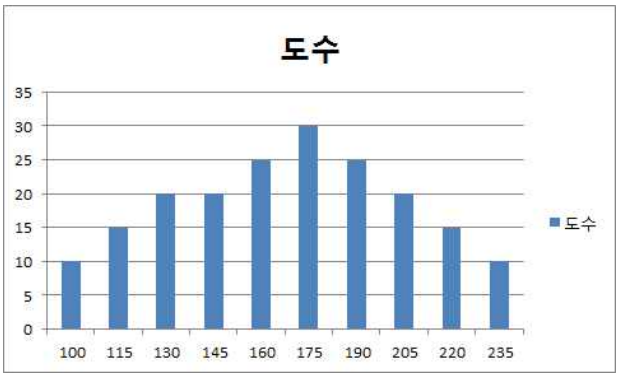
● 학생들의 몸무게가 제각각이라면



● 중앙으로 집중된 정도가 크다면



● 중앙으로 집중되었지만 조금 더 넓게 분포한다면



15. 사분위범위로 산포도 측정하는 방법

: 데이터를 4개의 동일 그룹으로 나눈 다음 양 끝은 그룹이 얼마나 멀리 떨어져 있는지를 알아보는 것

● 데이터를 숫자 순으로 정리

● 낮은 두개 그룹과 높은 두개 그룹으로 나눈다. (중앙값이 데이터 점이면 양쪽 모두에 포함시킨다)

● 낮은 그룹의 중앙점을 찾는다. 이것은 첫번째 사분위 Q_1 이 된다.

● 높은 그룹의 중앙값을 찾는다. 이것은 세번째 사분의 Q_3 가 된다.

● 사분위범위(IQR, Inter-Quartile Range)는 이들 사이의 거리이다.

$$IQR = Q_3 - Q_1$$

16. 펜실베니아 데이터로 IQR 찾아보기

- 데이터 정렬(줄기-잎 그림 활용), n=92

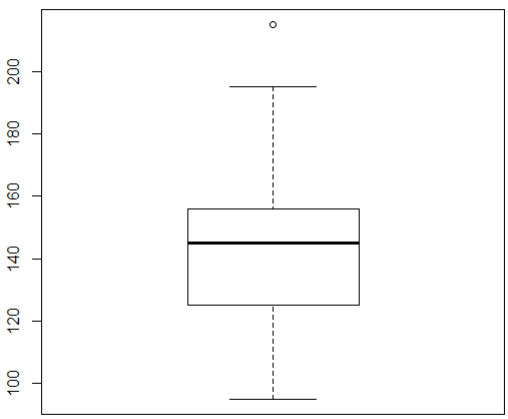
```
 9  : 5
10  : 288
11  : 002556688
12  : 000123555555
13  : 0000013555688
14  : 000025555 58
15  : 00000000003555555555557
16  : 000045
17  : 000055
18  : 0005
19  : 00005
20  :
21  : 5
```

- $\tilde{x}$ = (x_{46}+ x_{47})/2 = (145 +145)/2 = 145
- Q_1 = (x_{23}+x_{24})/2 = (125+125)/2 = 125
- Q_3 = (x_{69}+x_{70})/2 = (155+157)/2 = 156
- IQR = Q_3 − Q_1 = 156-125 = 31 파운드

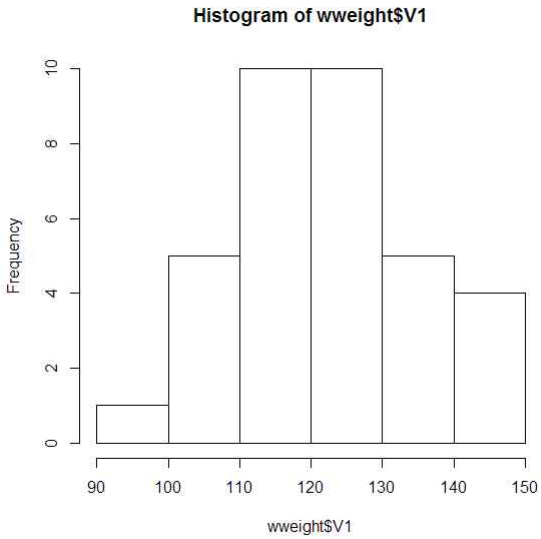
- 몸무게가 큰 학생들과 작은 학생들의 중앙값의 차이

17. Box Plot, 상자수염 그래프

- 상자의 양 끝은 Q_1, Q_3
- 중앙값은 상자안
- 상자의 끝에서 1.5 IQR 이상 떨어진 점은 이상(abnormal)값 (별도로 하나씩 따로 그림)
- 이상값이 아닌 가장 먼 점(1.5IQR 이내)까지 수염



18. 예제 : (펜실베니아 대학 여학생들 데이터만으로) 히스토그램 그려보기



19. 예제 : 줄기 잎 그림(stem-leaf)

```
 9 | 5
10 | 288
11 | 002556688
12 | 0001255555
13 | 0001358
14 | 05
15 | 000
```

20. 예제 : Box Plot