Lecture Notes for

Introduction to Probability and Statistics

Hanyang University, Ansan, 425-791, Korea

calltimy@hanyang.ac.kr

Probability and Statistics / 확률과 통계

강의노트 01

개요, 데이터 분석1

1. 데이터분석, 확률, 통계적 추론 - 통계학자들이 필요로 하는 세가지 근거

- 데이터분석 - 데이터의 수집, 전시, 요약
- 확률 - 카지노에서 비롯된 승산의 법칙
- 통계적 추론 - 확률 지식을 이용해 특정 데이터에서 통계적 결론을 이끌어내는 과학

2. 펜실베니아주 92명 학생의 몸무게 데이터 (Larry Gonick, The cartoon guide to statistics)

남학생

140 145 160 190 155 165 150 190 195 138
160 155 153 145 170 175 175 170 180 135
170 157 130 185 190 155 170 155 215 150
145 155 155 150 155 150 180 160 135 160
130 155 150 148 155 150 140 180 190 145
150 164 140 142 136 123 155

여학생

140 120 130 138 121 125 116 145 150 112
125 130 120 130 131 120 118 125 135 125
118 122 115 102 115 150 110 116 108 95
125 133 110 150 108

3. 확률, 통계 프로그램

SPSS , SAS , Matlab , Open Source Program R, Excel, …

4. 도수분포표

CASE I

계급	중앙값	도수	상대도수
87.5-102.4	95	2	.022
102.5-117.4	110	9	.098
117.5-132.4	125	19	.206
132.5-147.4	140	17	.185
147.5-162.4	155	27	.293
162.5-177.4	170	8	.087
177.5-192.4	185	8	.087
192.5-207.4	200	1	.011
207.5-222.4	215	1	.011
합계		92	1.000

CASE II

계급	중앙값	도수	상대도수
88-102	95	2	.022
103-117	110	9	.098
118-132	125	19	.206
133-147	140	17	.185
148-162	155	27	.293
163-177	170	8	.087
178-192	185	8	.087
193-207	200	1	.011
208-222	215	1	.011
합계		92	1.000

- 도수 : 구간 혹은 계급에 해당되는 데이터의 개수
- 상대도수 : 구간 혹은 계급에 해당되는 데이터의 개수를 전체 데이터의 개수로 나눈 값

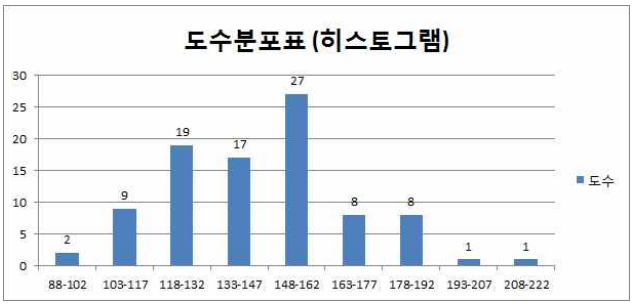
5. 계급을 정하는 방법

- 반올림된 숫자를 중앙값이 되도록 하고 같은 크기가 되도록 정한다.
- 데이터의 양이 적으면 계급의 개수도 적게, 양이 많으면 계급의 개수도 크게 한다.
- 율(Yule, G. U.)의 방법
  주관적인 결정방법으로 계급의 개수를 대략 15개에서 25개 정도로 한다.
- 스터지(Sturges, H. A.) - 전체 데이터의 수 N, 계급의 개수 K 일 때,
  $K=1+\log_2 N$

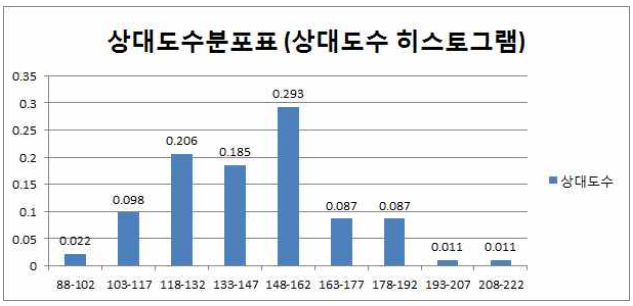
자료의 수	8	16	32	64	128	256
계급의 수	4	5	6	7	8	9

- 자료의 개수가 50~200개이면 $\sqrt{n}$ 을 중심으로 ±3의 범위에서 정한다.

6. 히스토그램 - 표로 되어 있는 도수 분포를 정보 그림으로 나타낸 것, 즉, 도수분포표를 그래프로 나타낸 것



히스토그램은 계급값 주변에 얼마나 많은 데이터들이 모여 있는지를 보여준다.



상대도수분포표는 도수분포표와 같은 모양을 지닌다. 수직축 값의 크기가 백분율로 표기된다.

7. 줄기-잎 그림

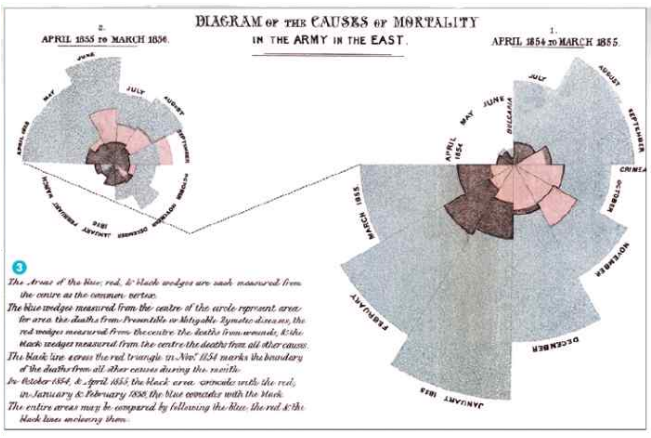
통계학자 존 터키가 고안한 각 데이터점을 그대로 유지하면서 요약하는 방법

```
 9  : 5
10  : 288
11  : 002556688
12  : 00012355555
13  : 0000013555688
14  : 00002555558
15  : 000000000003555555555557
16  : 000045
17  : 000055
18  : 0005
19  : 00005
20  :
21  : 5
```

- 줄기에는 몸무게 중 십자리 이상의 숫자만 쓴다.
- 잎에는 일의 자리 숫자를 쓴다.
- 모든 숫자를 다 쓴 후, 잎 부분의 숫자를 순서대로 정리한다.

8. 나이팅게일의 보고서 (Rose diagram, coxcomb)

- 크림전쟁 이후 영국군 병사들의 건강 상태에 대한 보고서에 실린 통계도표



보고서와 거기 실린 도표가 영국왕실의 마음을 움직였고, 42%에 달하던 병원내 사망률이 2%까지 떨어지게 되었다.

9. 대표값 과 산포도 - 측정값의 특징을 나타내는 중요한 두 가지 요소

10. 기호표시법 : x_1, x_2, x_3, … x_n

- n은 데이터의 전체 개수
- x4 는 데이터의 4번째 점의 값

측정	1	2	3	4	…	n
데이터값	x_1	x_2	x_3	x_4	…	x_n

- ex. 일주일 TV 시청 시간을 5명으로부터 조사한 자료

측정	1	2	3	4	5
데이터값	11	8	3	38	10

11. 대표값1 - 평균값(mean), $\bar{x}$

$$\bar{x} = \frac{\text{데이터의 합}}{n}$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$\because \sum_{i=1}^{n} x_i$: i 가 1부터 n 까지 일 때, x_i 의 합

펜실베니아주 92명 학생들의 몸무게의 평균

$$= \frac{\sum_{i=1}^{n} x_i}{n} = \frac{13{,}352}{92} = 145.15 \text{ 파운드}$$

12. 대표값2 - 중앙값(median), $\tilde{x}$

중앙값 : 데이터를 작은 것부터 순서대로 정렬 한 다음 그 한 중앙에 있는 값

5인의 TV 시청 시간 : 3, 8, **10**, 11, 38
중앙값 : 10

데이터의 개수가 짝수인 경우 : 중앙부분 2개의 값의 평균을 중앙값으로 한다.

4인의 TV 시청시간 : 3, 8, **10**, 11
중앙값 : (8+10)/2 = 9

펜실베니아주 92명 학생들의 몸무게의 중앙값

- 92는 짝수
- 46번째와 47번째의 몸무게의 평균이 중앙값

$$\tilde{x} = \frac{x_{46} + x_{47}}{2} = 145 \text{ 파운드}$$

13. 대표값이 하나가 아닌 이유(평균값, 중앙값, 기타)

- 평균값은 모든 데이터를 합친 값을 그 숫자만큼 나눈 값으로 기본적인 대표값으로 인정된다.
- 중앙값은 데이터들 중 다른 것과 차이가 극단적인 값에 민감하지 않기 때문에 데이터의 숫자가 크지 않은 경우 평균보다 오히려 나은 자료가 될 수 있다.
- 자료10의 TV 시청시간의 경우 만약 어떤 한 사람이 매주 200 시간을 본다면
- $\bar{x} = \frac{3+8+10+11+200}{5} = 46.4 \ hr$
- $\tilde{x} = \frac{10+11}{2} = 10.5 \ hr$