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

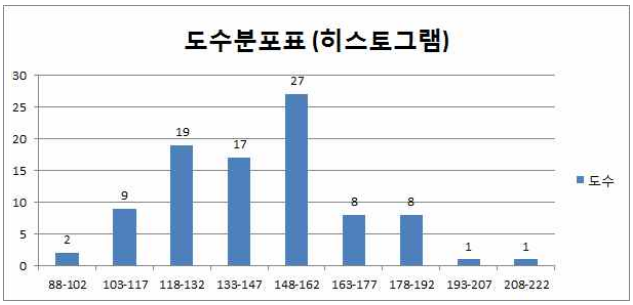
5. 계급을 정하는 방법

- 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,
  $K=1+\log_2 N$

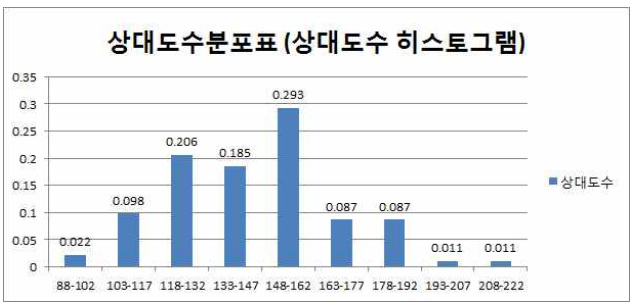
자료의 수	8	16	32	64	128	256
계급의 수	4	5	6	7	8	9

- 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

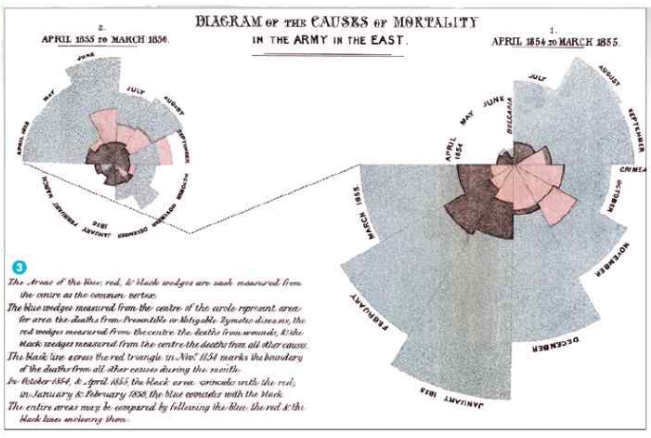
통계학자 존 터키가 고안한 각 데이터점을 그대로 유지하면서 요약하는 방법

```
 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

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

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
- $\bar{x} = \frac{3+8+10+11+200}{5} = 46.4 \ hr$
- $\tilde{x} = \frac{10+11}{2} = 10.5 \ hr$

Lecture Notes for
Introduction to Probability and Statistics
© 2014, Sungkuk Lim
Hanyang University, Ansan, 425-791, Korea
calltimy@hanyang.ac.kr

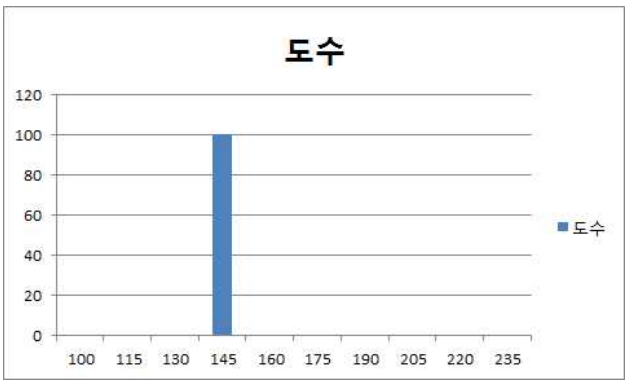
Probability and Statistics / 확률과 통계

강의노트 02

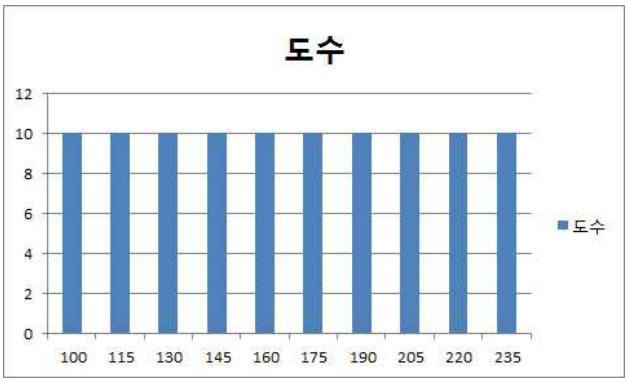
데이터분석 2

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

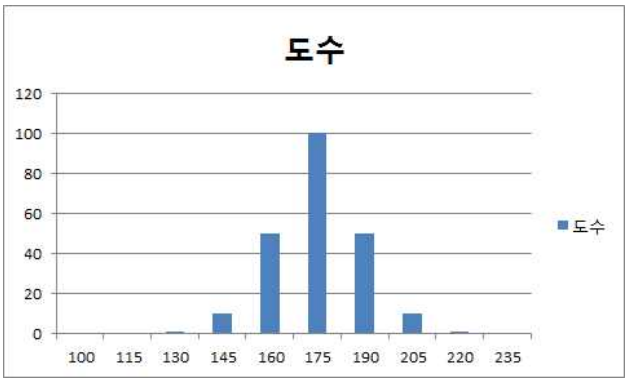
- 모든 학생의 몸무게가 145 파운드라면



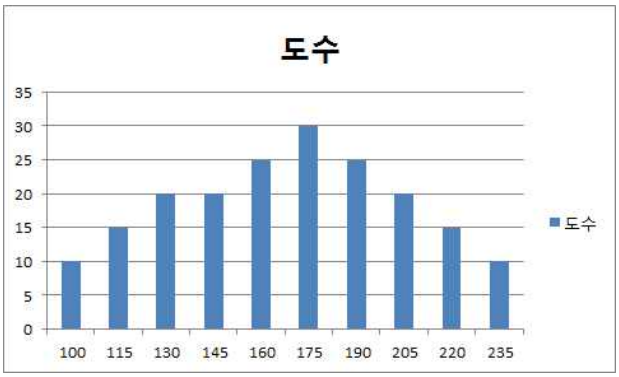
- 학생들의 몸무게가 제각각이라면



- 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

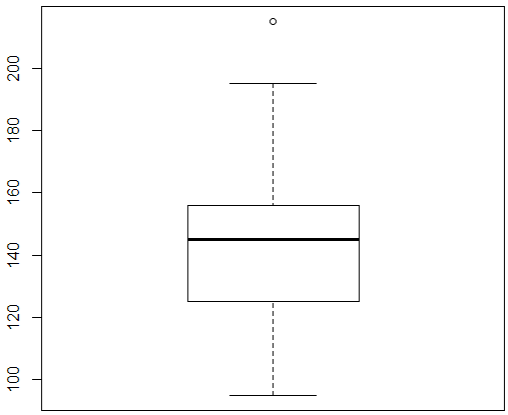
```
 9  : 5
10  : 288
11  : 002556688
12  : 000123555555
13  : 0000013555688
14  : 000025555588
15  : 00000000003555555555557
16  : 000045
17  : 000055
18  : 0005
19  : 00005
20  :
21  : 5
```

- $\tilde{x}$ = (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

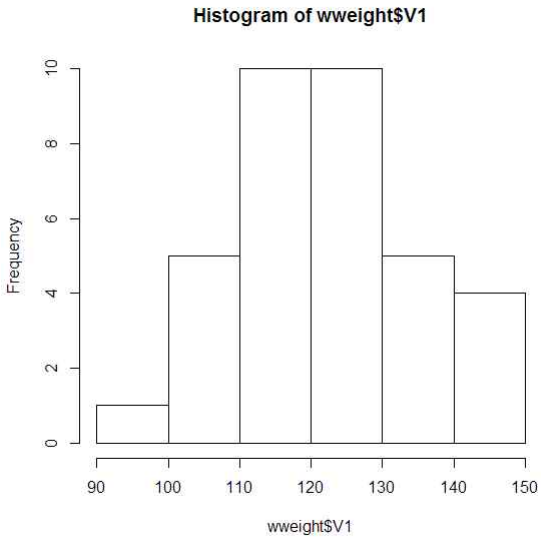
- 몸무게가 큰 학생들과 작은 학생들의 중앙값의 차이

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



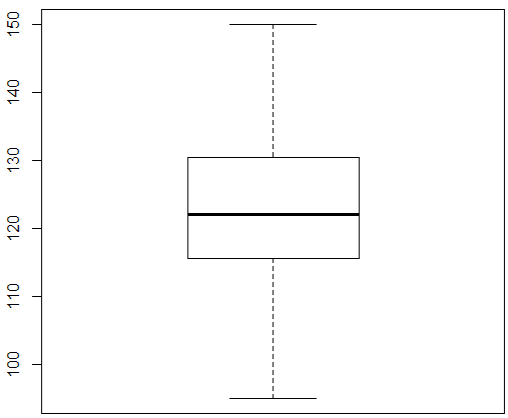
18. 예제 : (펜실베니아 대학 여학생들 데이터만으로) 히스토그램 그려보기



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



Lecture Notes for
Introduction to Probability and Statistics
© 2014, Sungkuk Lim
Hanyang University, Ansan, 425-791, Korea
calltimy@hanyang.ac.kr

Probability and Statistics / 확률과 통계

강의노트 03

확률 1

21. 가능성의 법칙인 확률의 시작은 도박이었다. 도박이 언제 시작되었는지는 분명하지 않다. 로마의 클라우디우스는 최초의 도박책을 쓴 것으로 전해진다. [주사위 놀이에서 이기는 법] 이라는 그의 책은 이름만 전해지고 있다.

22. 현재의 주사위는 르네상스 시기 도박사 드 메레가 수학적 문제를 제기하면서 일반화된다. 드 메레는 친구인 파스칼에게 주사위에 관한 수학적 질문을 하게 되고 파스칼은 현재 사용되는 확률 이론을 처음으로 만들게 된다.

23. 기본정의

- 확률실험, 시행 - 우연이 지배하는 사건의 결과를 관찰하는 과정
- 근원사건 - 어떤 시행에서 일어날 수 있는 모든 결과
- 표본공간 - 모든 근원사건의 집합

24. 동전던지기

- 시행 - 동전 던진 결과를 기록하는 것
- 근원사건 - 동전의 앞면, 뒷면
- 표본공간 - { 앞면, 뒷면 }

25. "1개의 주사위 던지기"

- 시행 - 주사위 던진 결과를 기록하는 것
- 근원사건 - 1, 2, 3, 4, 5, 6
- 표본공간 - {1, 2, 3, 4, 5, 6}

26, "2개의 주사위 던지기"

- 시행 - 주사위를 던져 나온 수를 합해서 기록하는 것
- 근원사건 - 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
- 표본공간 - 36개 (6x6) 로 이루어짐

{
(1,1), (1,2), (1,3), (1,4), (1,5), (1,6),
(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),
(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),
(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),
(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),
(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)
}

27. n 개의 결과를 가지는 확률시험에서 각각의 결과를 결과1, 결과2, 결과3, …, 결과N 이라고 쓰기로 하고, 결과1이 나올 확률을 P(결과1) 이라고 쓰기로 한다. 그러면 O_1, O_2, O_3,…, O_n 의 결과에 대응하는 확률은 $P(O_1)$, $P(O_2)$, $P(O_3)$,…, $P(O_n)$ 이 된다.

28. 2개의 주사위를 굴리는 경우, 36개의 근원사건이 있고, 모두 그 가능성이 같으므로 각 확률은 .0278 (=1/36) 이다.

29. 단, 위의 결과가 나오려면 주사위에서 각각의 눈이 나오는 확률이 같아야 한다는 전제가 따른다. 다르게 말하면 1의 값이 더 잘 나오도록 납을 넣어서 만든 주사위가 있을 수도 있다. 예를 들어 1은 굴린 회수의 25%가 나온다고 하면,

✓ 시행 – 주사위 던진 결과를 기록하는 것
✓ 근원사건 – 1, 2, 3, 4, 5, 6
✓ 표본공간 – {1, 2, 3, 4, 5, 6}

으로 25번의 결과와 동일하다. 하지만 정상적인 주사위라면 1부터 6까지 눈이 나올 확률은 1/6 이 되어야 하지만 이 주사위는 다음과 같다.

$P(O_1)$	$P(O_2)$	$P(O_3)$	$P(O_4)$	$P(O_5)$	$P(O_6)$
.25	.15	.15	.15	.15	.15

※ 근원사건 모두가 같은 확률을 가질 필요는 없고, 또 그런 경우는 일반적이지도 않다.
P(내일:비) ≠ P(내일:맑음) ≠ P(내일:구름)

30. 확률 접근 방법

- 고전적 확률 – 도박에 바탕을 둔 개념, 모든 근원사건은 동일한 확률을 가진다고 가정
- 통계적 확률 – 반복 가능한 시행에서 한 사건이 일어날 확률은 오래동안 관찰할 때 그 사건이 일어날 횟수의 비율
- 개인적, 주관적 확률 – 대부분의 사건은 일상생활에서 반복적으로 일어나지 않는다. 반복적이지 않은 어떤 일에 대해 개인이 평가하는 확률(ex, 북한이 핵실험을 재개할 확률, 대학생활중 CC가 될 확률, etc.)

31. 확률의 특징

- 확률은 0과 1사이에 있다.
- 확실히 일어나는 사건은 확률이 1이다.
- 확률은 음수가 될 수 없다.
- 확률 0 은 어떤 사건이 결코 일어날 수 없음을 의미한다.
- 모든 근원사건의 확률의 합은 1이다.

32. 고전적확률(Classical Formula)

- $P[A] = \frac{n(A)}{n(S)}$
- n(A) – 사건 A 가 일어날 경우의 수
- n(S) – 발생 가능한 모든 사건이 일어날 경우의 수

33. 통계적확률
(Relative Frequency Approximation)

- $P[A] \simeq \frac{f}{n}$
- f – 사건 A 가 발생하는 횟수
- n – 전체 실험(시행)의 횟수

34. 용어정리 1

- A', A^c : 사건 A의 여사건(complement): 사건 A에 포함되지 않은 표본공간에 속하는 모든 기본결과의 모임
- A∪B : 합사건(union) : 사건 A나 B 둘 중에 하나에 속하거나 동시에 속하는 기본결과들의 모임
- A∩B : 공통(또는 교)사건(intersection) : 사건 A와 B에 동시에 속하는 기본결과들의 모임
- ∅ : 공사건(null event, impossible event): 기본결과를 하나도 포함하지 않는 사건
- 상호배반(mutually exclusive) : 두 사건 A와 B의 교사건이 공사건
- **Two events A1 and A2 are mutually exclusive *if and only if* A1 ∩ A2 = ∅. Events A1, A2, A3,... are mutually exclusive *if and only if* Ai ∩ Aj = ∅ for i≠j**

35. 순열(Permutations) 과 조합(Combinations)

- 순열, nPr : 서로 다른 n개 중에서 순서를 생각하고 r개를 뽑는 경우의 수

  $${}_nP_r = \frac{n!}{(n-r)!}$$

- 조합, nCr : 서로 다른 n개 중에서 순서를 생각하지 않고 r개를 뽑는 경우의 수

  $${}_nC_r = C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

- 계승, 팩토리얼, !

  n! = n(n-1)(n-2)…3·2·1

- 0! = 1
- 중복순열 nΠr : 서로 다른 n개 중에서 순서를 생각하고 중복을 허용하여 r개를 뽑는 경우의 수

  $${}_n\Pi_r = n^r$$

- 중복조합 nHr : 서로 다른 n개 중에서 순서를 생각하지 않고 중복을 허용하여 r개를 뽑는 경우의 수

  $${}_nH_r = {}_{n+r-1}C_r$$

■ Example 1.3.7. 16개의 포트는 사용중(u) 이거나 사용하지 않지만 작동가능하거나(n), 전혀 작동하지 않는(i) 세가지 중 하나의 상태가 된다. 10 u, 4 n, 2 i 가 되는 경우의 수는 총 몇 가지 인가?

(a) $${}_{16}C_{10} \cdot {}_6C_4 \cdot {}_2C_2 = \frac{16!}{10!6!}\frac{6!}{4!2!}\frac{2!}{2!} = 120{,}120$$

■ 동전 1개를 100번 던질 때 나오는 면의 경우의 수는?

(a) $$\begin{aligned} {}_n\Pi_r &= n^r \\ &= 2^{100} \\ &= 1267650600228229401496703205376 \end{aligned}$$

■ 풀어볼 문제 Ch1. Exercise 5, 9, 28, 29, 32

Lecture Notes for
Introduction to Probability and Statistics
© 2014, Sungkuk Lim
Hanyang University, Ansan, 425-791, Korea
calltimy@hanyang.ac.kr

Probability and Statistics / 확률과 통계

강의노트 04

확률 2

36. 확률에 관한 기본 정리 1

P[S] = 1, S는 모든 근원사건(Sample Space)
P[∅] = 0, ∅는 공사건(impossible event)
P[A] ≧ 0, A는 어떤 사건(근원사건의 집합)
P[A'] = P[A^c] = 1-P[A], 여사건
A1, A2, A3 가 상호배반(mutually exclusive)이면, P[A1∪A2∪A3…] = P[A1] + P[A2] + P[A3] + …
P[A1∪A2] = P[A1] + P[A2] - P[A1∩A2]

37. Example 2.1.3. 해수에 들어있는 납과 수은 함유정도를 조사한 결과 다음 결과가 나왔다.

납 함유 : .32 -> P[A1]
수은함유 : .16 -> P[A2]
납 또는 수은 함유 : .38 -> P[A1∪A2]

납과 수은에 모두 오염된 정도(P[A1∩A2])는 얼마인가?

(a)
P[A1∪A2] = P[A1] + P[A2] - P[A1∩A2]
P[A1∩A2] = P[A1] + P[A2] - P[A1∪A2]
= .32 + .16 - .38
= .10

38. 기본연산

사건 : 근원사건의 집합, 예를 들면 2개의 주사위(하나는 희고, 하나는 검은 주사위)를 던져 눈의 합이 7이 되는 것 - 사건

사건	사건에 속하는 근원사건	확률
A:숫자합 3	{(1,2),(2,1)}	P[A]=2/36
B:숫자합 6	{(1,5),(2,4),(3,3),(4,2),(5,1)}	P[A]=5/36
C:흰눈 1	{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)}	P[A]=6/36
D:검은눈 7	∅	P[A]=0/36

39. 확률에 관한 기본 정리 2

근원사건 대신 사건을 사용하여 논리적 연산으로 사건들을 결합, 다른 사건을 만들 수 있다.
AND, OR, NOT

- E and F : 사건 E 와 F 가 둘 다 일어난다.
- E or F : 사건 E 또는 F 가 일어난다.
- not E : 사건 E 가 일어나지 않는다.

40. 덧셈정리
P[E∪F] = P[E] + P[F] - P[E∩F] (36번 동일)
P[E OR F] = P[E] + P[F] - P[E AND F]

※ 배반사건(mutually exclusive)의 경우
P[E OR F] = P[E] + P[F]

41. 뺄셈정리

P[E] = 1 - P[NOT E]

사건 E를 두 주사위 모두 1이 나오는 경우 이외의 사건으로 정의하면,

(a) P[E] = 1 - P[NOT E]
= 1 - 1/36
= 35/36

42. 조건부 확률

주사위 시행을 조금 바꿔서 희색 주사위를 던진 다음 검은색 주사위를 던진다. 두 주사위 눈을 합해 3이 될 확률(사건 A)은 얼마인가?

사건에 속하는 근원사건 : {(1,2), (2,1)}

(a) P[A] = 2/36

흰색 주사위가 1이 나왔다(사건C). 그때 사건 A 의 확률은?

주사위를 던지기 전에는 표본공간이 36(=6x6)개의 근원사건을 가지고 있었지만 사건 C가 일어났기 때문에 근원사건은 C로 인해 축소된 표본공간(6)에 속하게 된다.

- 축소된 표본공간 = {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)}
- 이중 사건A 에 해당되는 것은 {(1,2)} 하나 뿐

(a) P[A] = 1/6 ? wrong!!

P[A|C] : C사건이 이미 일어난 조건에서 A사건이 일어날 확률, "C가 주어졌을 때 A의 확률"

(a) P[A|C] = 1/6

$$P[A|C] = \frac{P[A \text{ AND } C]}{P[C]}$$

A가 일어나면 A가 일어난다는 것은 확실(당연)하다.

P[A|A] = 1

A와 C가 상호배반이면,

P[A|C] = 0

정리하면

P[A AND C] = P[A|C]·P[C]

A와 C를 바꾸면

P[C AND A] = P[C|A]·P[A]

P[C AND A] = P[A AND C] 이므로

P[A|C]·P[C] = P[C|A]·P[A]

43. 독립사건과 종속사건

- 독립사건 : 두 사건 E, F 가 하나의 사건이 일어나든 일어나지 않든 다른 사건이 일어날 확률에 영향을 주지 않으면 E 와 F는 서로 독립이다. 예를 들어 1개의 주사위를 굴리는 것은 다른 주사위를 굴리는 데 아무런 영향을 주지 않는다.
- 종속사건 : 두 사건 E, F 가 하나의 사건이 일어나는 것이 다른 사건이 일어날 확률에 영향을 준다면 E 와 F는 종속이다. 날씨와 에어컨 판매량과의 관계.

44. 특별 곱셈정리

사건 E 와 F 가 독립이면 다음의 특별 곱셈정리가 된다.

P[E AND F] = P[E]·P[F]

C : 흰색 주사위가 1이 나오는 사건
D : 검은색 주사위가 1이 나오는 사건

P[C|D] = P[C AND D] / P[D]
= (1/36)/(1/6)
= 1/6

A ; 두 주사위 눈의 합이 3이 되는 사건

P[A|C] = P[A AND C]/P[C]
= P[(1,2)]/P[C]
= (1/36) / (1/6)
= 1/6
≠ P[A], (P[A]=1/18)

P[A AND C] = 1/36
P[A] = 1/18
P[C] = 1/6

P[A AND C] ≠ P[A]·P[C]

∴ A 와 C 는 독립이 아니다.

45. 확률 법칙 요약 정리

덧셈정리
P[E OR F] = P[E] + P[F] - P[E AND F]

특별 덧셈 정리 : E, F 가 상호배반일 때
P[E OR F] = P[E] + P[F]

뺄셈정리
P[E] = 1 - P[NOT E]

곱셈정리
P[E AND F] = P[E|F]P[F]

특별 곱셈정리 : E 와 F가 독립일 때
P[E AND F] = P[E]P[F]

46. 드 메레의 문제

주사위 1개를 네 번 던져서 적어도 6이 한번 이상 나오는 확률과 2개의 주사위를 24번 던져서 둘 다 6이 나오는 확률 중 어느 것이 더 높은가?

Lecture Notes for
Introduction to Probability and Statistics
© 2014, Sungkuk Lim
Hanyang University, Ansan, 425-791, Korea
calltimy@hanyang.ac.kr

Probability and Statistics / 확률과 통계

강의노트 05

확률 3

47. 베이즈의 정리, 잘못된 양성반응 패러독스

Question :
인구 천명당 한명꼴로 걸리는 희귀병이 있다. 이 병에 걸렸는지를 판단하는 진단법이 있다. 병을 가진 사람이 이 진단법을 시행했을 때는 99%로 양성반응을 나타낸다. 반면 병을 가지지 않은 사람이 이 진단법에 따를 경우 2%만이 양성반응을 보인다. 만약 당신이 이 진단법에 따라 검사를 했고 그 결과로 양성반응이 나왔다면 여러분이 병에 걸려있을 확률은 얼마인가?

Answer :
사건A : 피검자가 병에 걸려 있다.
사건B : 피검자가 양성반응을 보인다.

P[A] = .001
P[B | A] = .99
P[B | not A] = .02
P[A | B] = ?

이 질병을 치료하는데는 심각한 부작용이 나타날 수 있다. 그래서 의사는 변호사와 함께 확률학자인 조 베이즈를 방문했다. 이 문제는 토머스 베이즈(조 베이즈의 선조, 1744~1809)가 최초로 증명했던 정리를 사용하여 풀어내었다.

조의 방법

표본 공간을 4개의 배반사건으로 나눈다.

	A	NOT A
B	A **AND** B	NOT A **AND** B
NOT B	A **AND** NOT B	NOT A **AND** NOT B

표에 있는 각 사건의 확률을 찾는다.

	A	NOT A	합계
B	P[A **AND** B]	P[NOT A **AND** B]	P[B]
NOT B	P[A **AND** NOT B]	P[NOT A **AND** NOT B]	P[NOT B]
합계	P[A]	P[NOT A]	1

계산하면,

P[A AND B]
= P[B|A]P[A]
= .99 x .001
= .00099

P[NOT A AND B]
= P[B|NOT A]P[NOT A]
= .02 x .999
= .01998

계산결과를 정리하면,

	A	NOT A	합계
B	.00099	.01998	.02098
NOT B	P[A **AND** NOT B]	P[NOT A **AND** NOT B]	P[NOT B]
합계	.001	.999	1

남은 부분을 구한다.

P[A **AND** NOT B] = .001 - .00099 = .00001
P[NOT A **AND** NOT B] = .999 - .01998 = .97902

최종적인 표

	A	NOT A	합계	
B	.00099	.01998	.02098	P[B]
NOT B	.00001	.97902	.97903	P[NOT B]
합계	.001	.999	1	
	P[A]	P[NOT A]		

위 표를 이용하면,

$$P[A|B] = \frac{P[A \text{ AND } B]}{P[B]}$$

$$=.0472$$

.0472 의 의미는?
-> 양성반응이 나온 사람 중 4.72%가 병에 걸려 있다. 즉, 1000명이 검사를 받았을 때 그중 21명이 양성반응이 나오고, 그 21명중 한명만이 병을 가지고 있다. - 잘못된 양성반응 패러독스

	감염자	비감염자	
양성반응	1	20	21
음성반응	0	979	979
	1	999	1000

결과:치료는 잘못된 부작용이 심각하므로 양성반응을 보인 사람들에 대해서 정밀 검사를 시행(의사)

베이즈의 정리

$$P[A|B] = \frac{P[B|A] \cdot P[A]}{P[B|A] \cdot P[A] + P[B|\text{NOT } A] \cdot P[\text{NOT } A]}$$

정리 :
1. 알고자 하는 것 : P[A|B]
2. 알고 있는 것 : P[A], P[B|A], P[B|NOT A]

48. 베이즈의 정리 일반식

$$P[A_j|B] = \frac{P[B|A_j] \cdot P[A_j]}{\sum_{i=1}^{n} P[B|A_i] \cdot P[A_i]}$$

49. 베이즈의 정리 예제

태아의 기형여부를 검사하려는 목적으로 이루어지는 트리플테스트는 산모의 혈액속에 있는 3가지(triple)의 표지물질(알파태아단백 (alpha-fetoprotein), 비포합성 에스트리올 (unconjugated estriol), 융모성 성선자극호르몬 (human chorionic gonadotropin))의 농도를 통해 다운증후군을 비롯한 태아 기형들을 간접적으로 테스트하는 방법이다.

통계적으로 1만명의 산모에서 태어난 태아 중 다운증후군인 태아는 13명이다. 태아가 다운증후군일 때 산모가 트리플테스트에 양성반응이 나올 확률은 70% 이고, 태아가 다운증후군이 아닐 때 양성반응이 나올 확률은 4.92% 이다. 만약 어떤 산모가 트리플테스트 결과 다운증후군 양성진단을 받았다면 그때 태아가 다운증후군일 확률은?

Answer :
P[A] : 태아가 다운증후군일 확률
P[B] : 트리플테스트에 양성일 확률
P[A|B] = ?

P[A] = .0013 (P[notA]=1-P[A]=.9987)
P[B|A] = .7
P[B| **not** A] = .0492

$$P[A|B] = \frac{P[B|A]P[A]}{P[B|A]P[A] + P[B|notA]P[notA]}$$

$$= \frac{.7 * .0013}{.7 * .0013 + .0492 * .9987}$$

$$= .00185$$

※ 1.8% 의 태아가 다운증후군을 가진다.