

General Statistical Concept

(일반적인 통계적 개념)

-School on Statistics for Astronomers-

2009. 4. 17

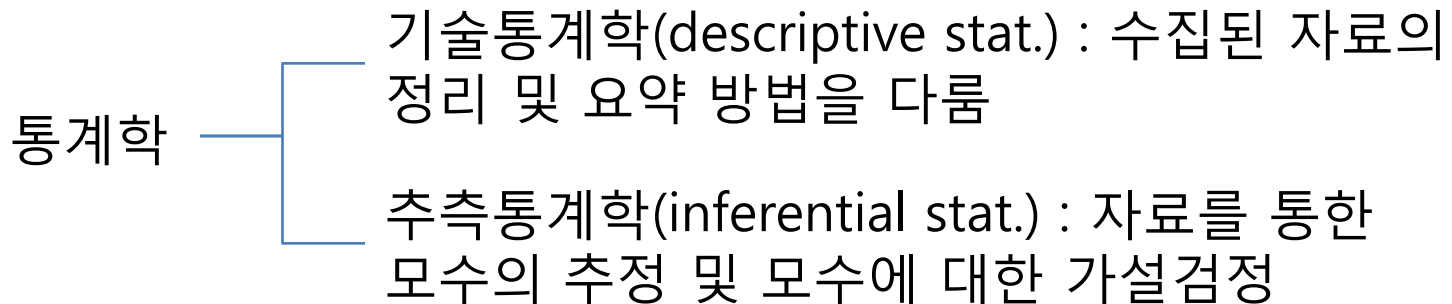
Honggie Kim

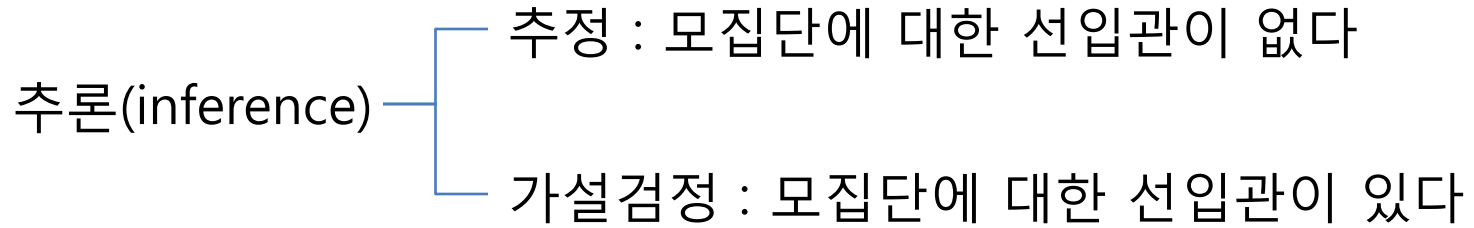
충남대학교 정보통계학과

통계학(Statistics)

관심의 대상에 대해 관련된 자료를 수집하고 그 자료를 요약, 정리하여 이로부터 불확실한 사실에 대한 결론이나 일반적인 규칙성을 추구하는 학문

Statistic : 통계치, 통계량





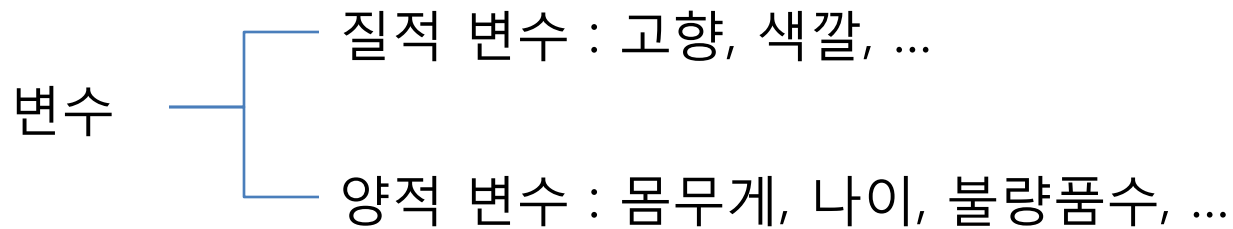
(예)

오늘 비가 오나 안오나 → 창문 열고 확인

9시 뉴스 일기예보(내일맑음) → 물소리

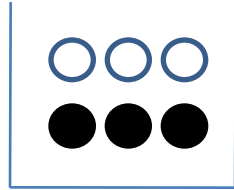
→ 비오나? → 창문 열고 확인

변수(variable) : 조사대상의 관심이 되는 특성



변수의 성질에 따라 통계적 분석 방법이 달라짐

확률 및 통계(probability and statistics)



흰공 3
검은공 3

두개를 추출



흰공 1, 검은공 1

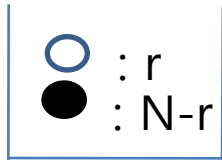
-비복원 추출(sampling without replacement)이면

$$P(\bullet \circ) = \frac{3}{6} \times \frac{3}{5} \times 2 = \frac{3}{5} \quad \text{or} \quad \frac{{}_3C_1 \times {}_3C_1}{{}_6C_2} = \frac{3 \times 3}{\frac{6 \times 5}{2}} = \frac{3}{5}$$

-복원 추출(sampling with replacement)이면

$$P(\bullet \circ) = \frac{3}{6} \times \frac{3}{6} \times 2 = \frac{1}{2} \quad \text{or} \quad {}_2C_1 \times \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^1$$

Prob model



total : N

n개를 추출 \longrightarrow X= 흰공의 갯수

-비복원 추출

$$P(X = x) = \frac{{}_r C_x \times {}_{N-r} C_{n-x}}{{}_N C_n} \longrightarrow \text{Hypergeometric}$$

-복원 추출

$$P(X = x) = {}_n C_x \left(\frac{r}{N}\right)^x \left(\frac{N-r}{N}\right)^{n-x}$$

$$= {}_n C_x p^x (1-p)^{n-x}, \quad p = \frac{r}{N} : \text{흰공의 비율} \longrightarrow \text{Binomial}$$

$$N \rightarrow \infty \longrightarrow \begin{cases} p \text{가 관심사} \\ \text{비복원, 복원 동일} \end{cases}$$

예 : 우리나라 국민 중 1000명을 뽑아 현 정부에 대한 지지여부를 조사



두개 비복원 추출



sample

알려진 사실 : $N=6$, 흰공과 검은공만 있다

흰공의 개수를 r 이라 하면

r	0	1	2	3	4	5	6
$P(\text{sample})$	0	$\frac{5}{15}$	$\frac{8}{15}$	$\frac{9}{15}$	$\frac{8}{15}$	$\frac{5}{15}$	0

$r=3$ 으로 추정

확률적 계산 : 알려진 모집단에서 주어진
표본이 얻어질 확률 계산

통계적 추론 : 주어진 표본으로부터 모집단에
대해 예측

모집단의 특성과 표본추출법에 따라
다양한 확률모형 존재

추정(Estimation)

(예) 친한 친구

상태	P(인상)	P(웃음)	합
배아프다	0.3	0.7	1
돈분실	0.8	0.2	1
실 연	0.2	0.8	1
기 타	0.01	0.99	1

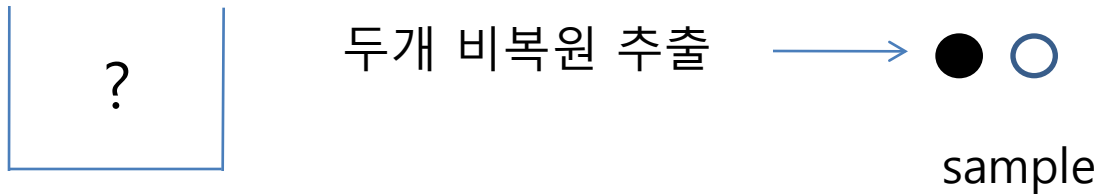
세 상태하에 각각의 확률 분포 존재

이 친구가 인상을 쓰고 있다면 어떤 상태일까?

- 돈분실 가능성이 가장 높다.

likelihood : 주어진 sample을 얻을 수 있는 확률을 모수의 함수로 본 것으로 확률과 같은 개념이나 모든 모수에 대한 합이 1이 아님.

maximum likelihood estimation : likelihood값이 최대가 되는 모수값을 추정치로 채택



알려진 사실 : N=6, 흰공과 검은공만 있다

흰공의 개수를 r이라 하면


r	0	1	2	3	4	5	6
P(sample)	0	$\frac{5}{15}$	$\frac{8}{15}$	$\frac{9}{15}$	$\frac{8}{15}$	$\frac{5}{15}$	0

r=3으로 추정

(예) 친한 친구(계속)

	상태	P(인상)	P(상태 ∩ 인상)
0.5	배아프다	0.3	$0.5 \times 0.3 = 0.150$
0.09	돈분실	0.8	$0.09 \times 0.8 = 0.072$
0.01	실 연	0.2	$0.01 \times 0.2 = 0.002$
0.4	기 타	0.01	$0.4 \times 0.01 = 0.04$
			0.264

인상  "배아프다"고 추정

모수가 동일한 가능성을 가지 않는다는 전제하에 추정  Bayesian 추정

Bayes 정리(혹은 공식)

A1, A2, ..., An이 S를 분할할 때,

$$P(A_i / B) = \frac{P(A_i) \cdot P(B / A_i)}{\sum P(A_i) \cdot P(B / A_i)}$$

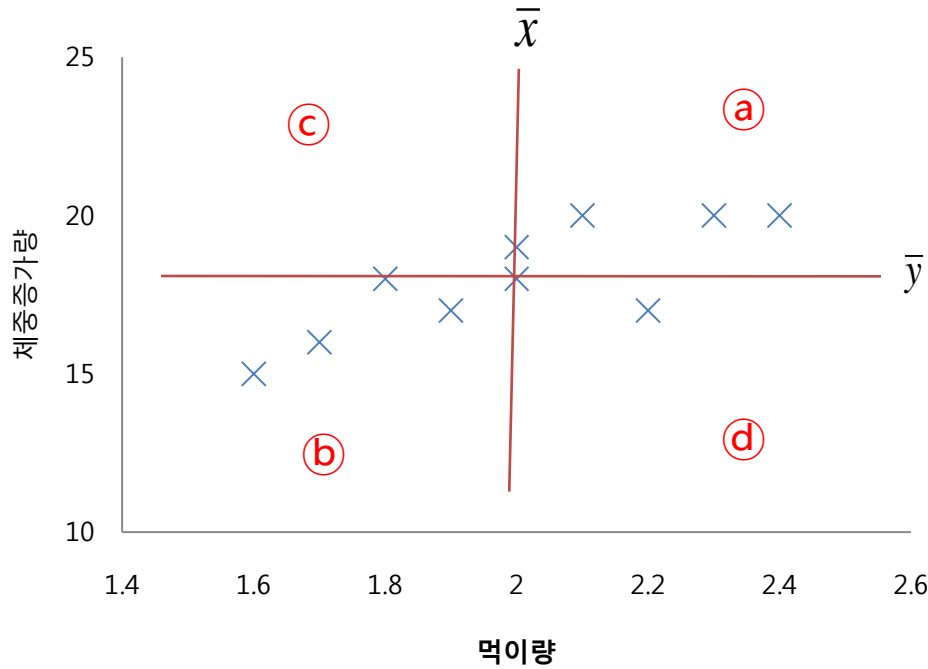
상태	사전확률	인상	사후확률
배아프다	0.50	→	0.568 = $\frac{0.150}{0.264}$
돈분실	0.09		0.273
실연	0.01		0.008
기타	0.40		0.152
	1.00		1.001

상관계수와 인과관계(correlation and causation)

(예) 10마리의 돼지의 하루평균 먹이량과 한달간의 체중 증가량

	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
	먹이량(kg)	체중증가량(kg)					
	2.0	18	0.0	0	0	0	0
	1.7	16	-0.3	-2	0.6	0.09	4
	2.2	17	0.2	-1	-0.2	0.04	1
	1.6	15	-0.4	-3	1.2	0.16	9
	2.4	20	0.4	2	0.8	0.16	4
	2.0	19	0.0	1	0	0.0	1
	1.8	18	-0.2	0	0	0.04	0
	2.1	20	0.1	2	0.2	0.01	4
	2.3	20	0.3	2	0.6	0.09	4
	1.9	17	-0.1	-1	0.1	0.01	1
sum	20	180	0	0	3.3	0.30	28
평균	$2.0 = \bar{x}$	$18 = \bar{y}$					

산점도



ⓐ, ⓑ 영역에서는 $(x_i - \bar{x})(y_i - \bar{y}) > 0$

ⓒ, ⓓ " " < 0

$\sum (x_i - \bar{x})(y_i - \bar{y}) > 0 \iff$ 양의 상관관계

$\sum (x_i - \bar{x})(y_i - \bar{y}) = 0 \iff$ 무상관

$\sum (x_i - \bar{x})(y_i - \bar{y}) < 0 \iff$ 음의 상관관계

$$\begin{aligned}
 & \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} \\
 r = & \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}} \\
 = & \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \\
 = & \frac{3.3}{\sqrt{0.6} \sqrt{2.8}} \\
 = & 0.805
 \end{aligned}$$

상관계수 : 두 변수간의 선형관계의 강도를 -1과 1사이의 수치로 표현

(예)

아빠와 노는 시간이 많았던 아이들의 IQ가 높다는 연구결과

 \Rightarrow 아빠가 많이 데리고 놀면 IQ가 높아질까?

(예) 키와 몸무게

 \Rightarrow 키가 커지면 몸무게가 증가
Causation 존재 $\Rightarrow \hat{y} = a + bx$ (회귀직선의 식)

$$\begin{aligned}\hat{w} &= (h - 100) \times 0.9 \\ &= -90 - 0.9h\end{aligned}$$

 $h = 174$ 이면 $\hat{w} = 66.6$ 이 정상체중

가설검정(hypothesis test)

(예) 9시 수업이 있는데 늦잠을 잤다.

- 택시를 타고 수업에 간다.
- 버스를 타고 가다 늦으면 제깬다.

평범한 날			시험보는 날		
Action	cost	reward	Action	cost	reward
택시	돈	수업	택시	돈	학점
버스	수업	돈	버스	학점	돈

⇒ Cost가 보편적인 판단기준

두 종류의 error

Type I error = H_0 가 사실인데 기각

Type II error = H_0 가 거짓인데 불기각

H0	action	
	reject	do not reject
True	Type I error	OK
false	OK	Type II error

일반적으로 type I error가 더 심각

$$P(\text{type I error}) = \alpha$$

$$P(\text{type II error}) = \beta$$

α 와 β 가 모두 적은 결정 법칙

α 가 감소하면 β 가 증가

H_0 : 평범한 땅

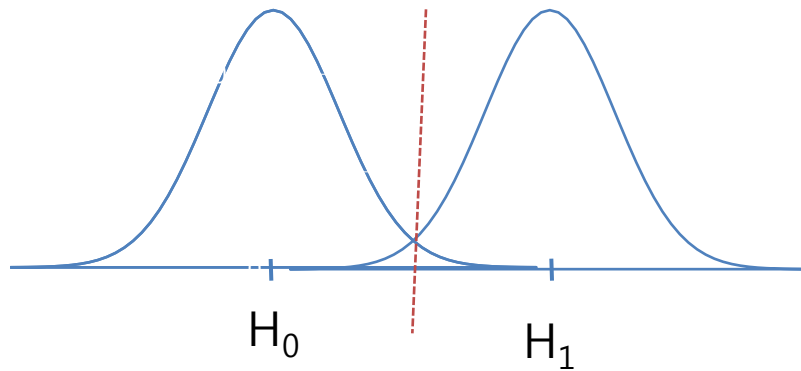


▲	돌도끼	한 개	발견
▲▲	"	두 개	"
⋮	⋮	⋮	⋮

돌도끼 100개, 조개껍질무덤, ...



H_1 : 선사유적지



고고학회
 H_1 : 선사유적지
 H_0 : 평범한 땅

(예) 사법 시험

 H_0 : 평범한 사람 H_1 : 범조인 $\alpha = P(\text{평범한 사람을 범조인으로})$ $\beta = P(\text{범조인 자격이 있는 사람을 평범한 사람으로})$

자격이 있는 사람이 떨어지면 내년 재시험

사법시험은 α 와 β 가 무지 작은 시험

위 재판

 H_0 : 무죄 H_1 : 유죄 $\alpha = 0$, β 가 큼.

(예) 동전의 앞면 확률 p

$$H_0 : p=0.5$$

$$H_1 : p=0.8$$

test : 15번 던져서 몇 번 이상 나오면 H_0 를 버릴 것인가?

α 가 주어져야 결정가능!!

$$X \sim \text{Binomial}(n, p)$$

$$P(X \geq x / p = 0.5) \approx 0.05 \quad \text{인 } x \text{를 찾으면}$$

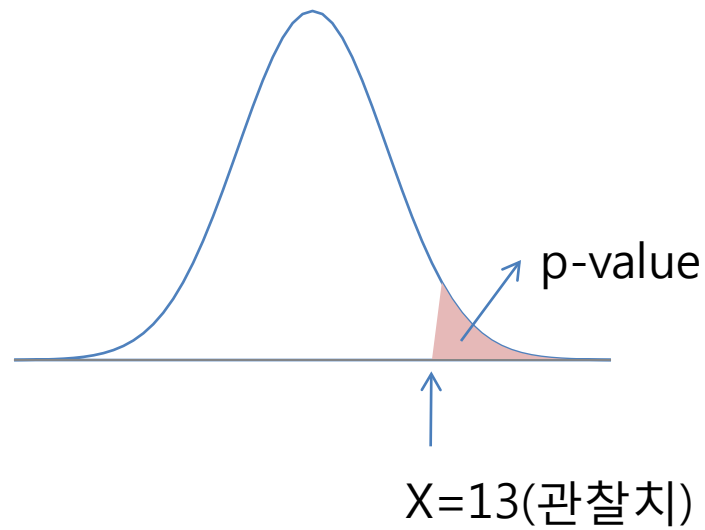
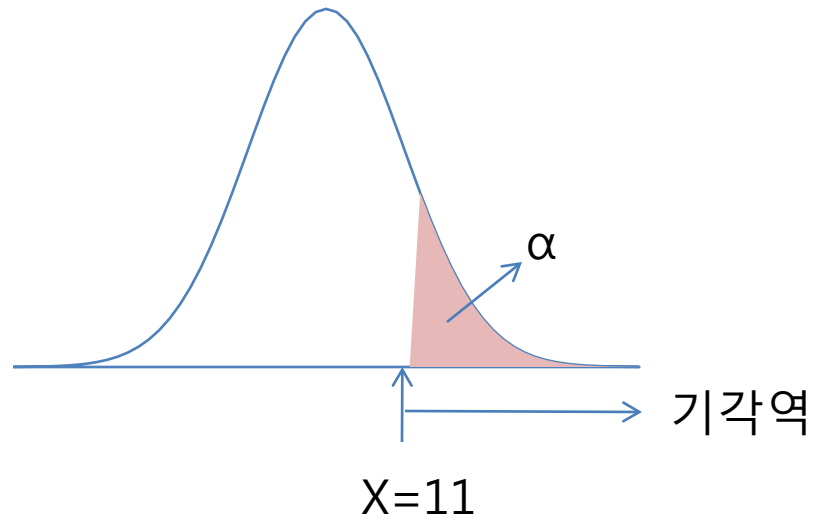
$$X \geq 11 \text{ 이면 } H_0 \text{ 기각, 이때 } \alpha = 0.059$$

$$\beta = P(X \leq 10 / p = 0.8) = 0.061$$

$$P(X \leq 4 / p = 0.5) = 0.059$$

$$X \leq 4 \text{ 이면 } H_0 \text{ 기각도 } \alpha = 0.059$$

$$\text{하지만 } \beta = P(X \geq 5 / p = 0.8) = 1.0$$



p-value(유의확률): H_0 가 참일 때 검정통계량이 표본에서 계산된 값과 같거나 그 값보다 대립가설 방향으로 더 극단적인 값을 가질 확률

컴퓨터 통계패키지는 p-value 제공

$p\text{-value} \leq \alpha \Rightarrow$ reject H_0

$p\text{-value} > \alpha \Rightarrow$ do not reject H_0