

Statistical Inference

Estimation and Testing Hypothesis

김 주 한

충남대학교

자연과학대학 정보통계학과

1. 개체 n 개의 특성 k 가지를 관측하여 얻은 데이터
표본수가 n 이고 변수가 k 개인 데이터

	변수 1	변수 2	...	변수 k
개체 1	x_{11}	x_{21}	...	x_{k1}
개체 2	x_{12}	x_{22}	...	x_{k2}
⋮	⋮	⋮		⋮
개체 n	x_{1n}	x_{2n}	...	x_{kn}

2. 데이터에 대한 통계적 모형

표본수가 n 이고 변수가 k 개인 데이터 $(x_{11}, \dots, x_{1n}), (x_{21}, \dots, x_{2n}), \dots,$
 (x_{k1}, \dots, x_{kn}) 은 확률변수 $(X_{11}, \dots, X_{1n}), (X_{21}, \dots, X_{2n}), \dots, (X_{k1}, \dots, X_{kn})$ 의
관측값이다.

- (1) 일변량(univariate) 모형

X_{i1}, \dots, X_{in} 은 서로 독립이고 확률분포 $f_{ij}(x)$ ($j=1, \dots, n$)를 따른다.

(예) X_{i1}, \dots, X_{in} 은 $N(\mu, \sigma^2)$ 확률표본이다. 즉, X_{i1}, \dots, X_{in} 은 서로
독립이고 평균이 μ 분산이 σ^2 인 정규분포를 따르는 확률변수이다.

- (2) 다변량(multivariate) 모형

$(X_{11}, \dots, X_{k1}), (X_{12}, \dots, X_{k2}), \dots, (X_{1n}, \dots, X_{kn})$ 은 서로 독립이고
확률분포 $f_j(x_1, \dots, x_k)$ ($j=1, \dots, n$) 를 따른다.

(예) $(X_{11}, X_{21}), (X_{12}, X_{22}), \dots, (X_{1n}, X_{2n})$ 은 서로 독립이고
평균이 μ_1, μ_2 , 분산이 σ_1^2, σ_2^2 , 상관계수가 ρ 인
이변량 정규분포를 따른다.

(3) 모수적(parametric) 모형

확률분포 $f_{ij}(x)$ ($j=1, \dots, n$) 의 형태가 모수(parameter)에 의해 결정되는 확률모형.

(예) $N(\mu, \sigma^2)$ 확률표본

(4) 비모수적(nonparametric) 모형

확률분포 $f_{ij}(x)$ ($j=1, \dots, n$) 의 형태를 모르는 경우.

(예) X_{i1}, \dots, X_{in} 은 서로 독립이고 평균이 μ 분산이 σ^2 인 확률변수이다.

3. 확률변수에 대한 확률모형 : 확률분포

(1) 이산형 일변량 확률분포

이항(binomial)분포, 포아송(Poisson)분포,
기하(geometric)분포, 음이항(negative binomial)분포

(2) 연속형 일변량 확률분포

정규(normal)분포, 지수(exponential)분포,
감마(gamma)분포, 베타(beta)분포

(3) 이산형 다변량 확률분포

다항(multinomial)분포

(4) 연속형 다변량 확률분포

다변량정규(multivariate normal)분포

4. 임의로 뽑은 충남대학교 학생 data

	성별	신장	체중
1	0 (남) x_1	180 y_1	85 z_1
2	0 (남) x_2	168 y_2	70 z_2
3	0 (남) x_3	174 y_3	72 z_3
4	1 (여) x_4	161 y_4	55 z_4
5	1 (여) x_5	164 y_5	56 z_5

(x_i, y_i, z_i) , $i = 1, 2, \dots, 5$ 는 확률변수 (X_i, Y_i, Z_i) 의 관측값

X_i : i 번째 표본의 성별, 이산형 확률변수

Y_i : i 번째 표본의 신장, 연속형 확률변수

Z_i : i 번째 표본의 체중, 연속형 확률변수

일변량 통계 모형

- (1) X_i 는 모수가 $n=1$ 과 p 인 이항분포를 따른다.
- (2) Y_i 는 평균이 μ_y 이고 분산이 σ_y^2 인 정규분포를 따른다.
- (3) Z_i 는 평균이 μ_z 이고 분산이 σ_z^2 인 정규분포를 따른다.

다변량 통계모형

- (1) $X_i = 0$ 이면 (Y_i, Z_i) 는 모수가 $(\mu_{y0}, \mu_{z0}, \sigma_{y0}^2, \sigma_{z0}^2, \rho_0)$ 인 이변량 정규분포를 따른다.
- (2) $X_i = 1$ 이면 (Y_i, Z_i) 는 모수가 $(\mu_{y1}, \mu_{z1}, \sigma_{y1}^2, \sigma_{z1}^2, \rho_1)$ 인 이변량 정규분포를 따른다.

(5) 추정(estimation)

(a) Point estimation(점추정)

모수의 값을 추정

대전시에 거주하는 40대 남성의 혈중 콜레스테롤 레벨의 평균 μ 의 추정치는 185 이다.

(b) Interval estimation(구간추정) (신뢰구간)

95% 신뢰수준에서 $175 < \mu < 197$ 으로 추정

(6) 가설검증(Hypothesis testing)

모수에 대한 가설을 세우고 그 가설을 검증한다.

대전시에 거주하는 40대 남성의 혈중 콜레스테롤 레벨의 평균 μ 가 180 보다 크다고 할 수 있는지를 확인하려면 두 가설

$$H_0 : \mu \leq 180, \quad H_1 : \mu > 180$$

중 하나를 통계적 방법을 이용하여 선택한다.

6. 추론절차

- (1) 관심있는 모수에 대한 모든 정보를 포함하고 있는 통계량(statistic)을 찾는다. (최소충분통계량, minimal sufficient statistic)

X_1, X_2, \dots, X_n 이 $N(\mu, \sigma^2)$ 을 따르는 확률표본이면

μ 에 대한 최소충분통계량은 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 이다.

(2) 통계량의 확률분포를 구한다. (표본분포)

\bar{X} 의 확률분포는 $N\left(\mu, \frac{\sigma^2}{n}\right)$ 이다.

(3) 적당한 추정방법을 사용하여 모수에 대한 좋은 추정량을 찾는다.

좋은 추정량은 최소충분통계량의 함수이다.

- (a) 최대가능도추정량(Maximum Likelihood Estimator)
- (b) 적률추정량(Moment Estimator)
- (c) 최소분산불편추정량(Minimum Variance Unbiased Estimator)
- (d) 최소제곱추정량(Least Squares Estimator)
- (e) 최소카이제곱추정량(Minimum Chi-square Estimator)
- (f) 베이스추정량(Bayesian Estimator)

(4) 추정량의 좋은 성질

(a) 불편성(unbiasedness) (불편추정량)

$$\text{bias} = E(\hat{\theta}) - \theta = \text{추정량의 기대값} - \text{모수} = 0$$

(b) 작은 MSE(mean squared error)

$$\text{MSE} = E((\hat{\theta} - \theta)^2) = \text{추정량의 분산} + \text{bias}^2$$

(최소분산불편추정량 (MVUE))

(c) 일치성(consistency)

표본의 수가 커지면 추정량은 모수의 참값에 확률적으로 수렴한다.

(d) 점근적 정규성(asymptotic normality)

$\sqrt{n}(\hat{\theta} - \theta)$ 의 극한분포가 정규분포이다.

Best Asymptotically Normal estimator (MLE, MVUE)

(5) 최대가능도추정량(MLE)

확률변수 X 의 확률밀도함수(연속형) 또는 확률질량함수(이산형)를 $f(x|\theta)$ 라 하자. 확률변수 X 를 관측하여 얻은 관측값 $X = x_0$ 을 대입하면 $f(x_0|\theta)$ 는 θ 의 함수가 되고 이것을 $X = x_0$ 일 때 θ 의 가능도함수(likelihood function)라 하며 $L(\theta|x_0)$ 로 나타낸다. 임의의 관측값 x 에 대한 θ 의 가능도함수는 $L(\theta|x) = f(x|\theta)$ 이고 이 함수를 최대로 하는 θ 를 최대가능도추정량(MLE)이라 한다. 즉,

$$\text{MLE} : \hat{\theta}, \quad L(\hat{\theta}|x) = \max_{\theta} L(\theta|x)$$

(a) $X \sim B(4, \theta)$ (이항분포), 모수공간 : $\Theta = \{0.2, 0.5, 0.7, 0.9\}$

	x					
θ	0	1	2	3	4	
0.2	0.4096	0.4096	0.1536	0.0256	0.0016	1.0
0.5	0.0625	0.2500	0.3750	0.2500	0.0625	1.0
0.7	0.0081	0.0756	0.2646	0.4116	0.2401	1.0
0.9	0.0001	0.0036	0.0486	0.2916	0.6561	1.0

$x = 1$ 일 때 $\theta = 0.2$ 일 가능도는 0.4096이고

$\theta = 0.7$ 일 가능도는 0.0756이다

관측값 x	0	1	2	3	4
MLE $\hat{\theta}$	0.2	0.2	0.5	0.7	0.9

(b) $X \sim B(n, \pi)$, $0 \leq \pi \leq 1$ 이면 가능도함수는

$$L(\pi|x) = f(x|\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

이고, 로그가능도함수는

$$l(\pi|x) = \log L(\pi|x) = \log \binom{n}{x} + x \log \pi + (n-x) \log(1-\pi)$$

이다. 로그가능도함수를 미분하여 0으로 놓은 식을 가능도 방정식이라고 하고 이 가능도 방정식의 해가 π 의 MLE가 된다.

$$\frac{\partial l(\pi|x)}{\partial \pi} = \frac{x}{\pi} - \frac{n-x}{1-\pi} = 0$$

의 해는 $\pi = \frac{x}{n}$ 이므로 π 의 MLE는 $\hat{\pi} = p = \frac{X}{n}$ 이다.

$\hat{\pi} = p = \frac{X}{n}$ 를 표본비율이라 한다.

(c) 지수분포 예

n 개 개체의 수명 T_1, T_2, \dots, T_n 은 모수가 λ 인 지수분포를 따른다.

$$f(t_i|\lambda) = \lambda \exp(-\lambda t_i), \quad t_i > 0, \lambda > 0$$

$$f(t_1, \dots, t_n|\lambda) = f(t_1|\lambda) \cdots f(t_n|\lambda) = \lambda^n \exp(-\lambda \sum_{i=1}^n t_i) = \lambda^n \exp(-\lambda s)$$

$$\Rightarrow L(\lambda) = \lambda^n \exp(-\lambda s), \quad s = \sum_{i=1}^n t_i$$

$$\Rightarrow l(\lambda) = \log L(\lambda) = n \log \lambda - \lambda s$$

$$\Rightarrow \frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - s = 0 \Rightarrow \lambda = \frac{n}{s}$$

그러므로 λ 의 MLE는 $\hat{\lambda} = \frac{n}{\sum_{i=1}^n T_i} = \frac{1}{T}$

(6) 구간추정(interval estimation)

(a) 모수 θ 에 대한 $100(1-\alpha)\%$ 신뢰구간

두 통계량 L, U 에 대하여 $P(L < \theta < U) = 1 - \alpha$ 를 만족시키는 확률구간 $L < \theta < U$ 를 θ 에 대한 $100(1-\alpha)\%$ 신뢰구간이라 한다.

신뢰계수 : $1 - \alpha$

신뢰구간 길이 : $E(U - L)$ 또는 $U - L$ 의 계산된 값

신뢰구간의 길이는 신뢰계수와 표본의 크기에 따라 변한다.

(b) 신뢰구간 구하는 방법

모수 θ 의 좋은 추정량 $\hat{\theta}$ 의 표본분포를 이용한다. θ 와 $\hat{\theta}$ 의 함수이고 표준형 분포를 따르는 또는 근사적으로 표준형 분포를 따르는 확률변수 $Q(\hat{\theta}, \theta)$ 를 찾는다. 이러한 확률변수 $Q(\hat{\theta}, \theta)$ 를 피벗(pivot) 이라고 한다. $P(c_1 \leq Q(\hat{\theta}, \theta) \leq c_2) = 1 - \alpha$ 를 만족시키는 상수 c_1, c_2 를 찾은 후 부등식 $c_1 \leq Q(\hat{\theta}, \theta) \leq c_2$ 를 θ 에 대한 부등식으로 바꾼다.

(예) $Y \sim N(\mu, 4)$ 일 때 μ 에 대한 95% 신뢰구간

$$Q = \frac{Y - \mu}{2} \sim N(0, 1) : \text{피벗}$$

$$P(-1.96 \leq Q \leq 1.96) = 0.95$$

$$-1.96 \leq \frac{Y - \mu}{2} \leq 1.96 \Rightarrow Y - 3.92 \leq \mu \leq Y + 3.92$$

(c) 가설검증의 귀무가설 채택역을 이용한다.

(7) 가설검증(Hypothesis testing)

모집단의 분포나 모수에 대한 가설을 세우고, 모집단에서 추출한 표본에 기초하여 그 가설을 검증하는 통계적기법

(a) 대립가설(대안가설) (alternative hypothesis) (H_1)

효과가 있다, 차이가 있다, 서로 다르다 와 같은 형태의 가설로서 표본으로부터 확실한 근거를 찾아 입증하고자 하는 것

(예) 어느 국회의원 후보의 지지율은 30% 미만이다

$$H_1: \theta < 0.3, \theta: \text{국회의원 후보의 지지율}$$

(b) 귀무가설(영가설) (null hypothesis) (H_0)

효과가 없다, 차이가 없다, 서로 다르지 않다 와 같은 형태의 가설로 대립가설이 옳다는 강력한 증거를 찾지 못할 때 받아들이는 것이다.

(예) 국회의원 후보의 지지율은 30% 보다 작지 않다.

$$H_0: \theta \geq 0.3 \text{ 또는 } H_0: \theta = 0.3$$

(c) 검증통계량 (test statistics)

가설검증에 사용되는 통계량

(예) 어떤 국회의원 후보 지지율을 알아보기 위해 유권자

100명을 임의로 추출하여 국회의원 후보를 지지하는지 조사하였다. Y 를 100명 중 그 국회의원 후보를 지지하는 유권자의 수라 하면 Y 를 검증통계량으로 사용할 수 있다.

(d) 기각역 (rejection region)

주어진 유의수준에서 귀무가설을 기각하는 검증통계량의 값의 범위
검증통계량의 값이 기각역에 들어가면 귀무가설을 기각한다

(예) $Y \leq 22$ 이면 귀무가설을 기각한다.

(e) 제1종 오류 (type I error)

귀무가설이 사실일 때 귀무가설을 기각하는 오류

(예) 제1종 오류 확률

$$P(Y \leq 22 | \theta \geq 0.3) = \sum_{y=0}^{22} f(y|\theta), \theta \geq 0.3$$

$$f(y|\theta) = \binom{100}{y} \theta^y (1-\theta)^{100-y} : B(100, \theta) \text{ 확률함수}$$

(f) 제2종 오류 (type II error)

대립가설이 사실일 때 귀무가설을 채택하는 오류

(예) 제2종 오류 확률

$$P(Y > 22 | \theta < 0.3) = \sum_{y=23}^{100} f(y|\theta), \theta < 0.3$$

$$\theta = 0.25 (0.714), \theta = 0.20 (0.261), \theta = 0.15 (0.022)$$

(g) 유의수준 (significance level) α

제1종 오류 확률의 최대 허용한계

$$P(Y \leq 22 | \theta = 0.3) = \sum_{y=0}^{22} f(y|0.3) = 0.048 \leq \alpha$$

(h) 검증력 (power)

대립가설이 사실일 때 귀무가설을 기각하는 확률

검증력 = 1 - 제2종 오류 확률

(예) 검증력

$$P(Y \leq 22 | \theta < 0.3) = \sum_{y=0}^{22} f(y|\theta), \theta < 0.3$$

$$\theta = 0.25 (0.286), \theta = 0.20 (0.739), \theta = 0.15 (0.978)$$

(i) 최강 검증력 검증법 (most powerful test)

표본의 크기가 정해져 있을 때 제1종 오류 확률과 제2종 오류 확률 두 가지를 동시에 작게 할 수 없으므로 주어진 유의수준에서 검증력을 최대로 하는 (제2종 오류 확률을 최소로 하는) 검증방법을 사용한다.

(j) p 값, 유의확률 (p-value, significance probability)

귀무가설이 사실일 때, 검증통계량이 실제로 관측된 값과 같거나 그보다 대립가설방향으로 더 극단적인 값이 될 확률로 p 값이 주어진 유의수준보다 작으면 귀무가설을 기각한다.

(k) 한쪽검증 (one-sided test), 양쪽검증 (two-sided test)

대립가설에서 모수의 영역이 한 쪽으로만 주어지면 한쪽(단측)검증 양쪽에 주어지면 양쪽(양측)검증이라 한다.

(l) 한쪽검증의 p 값

검증통계량은 W 이고 w 는 표본으로부터 계산한 W 의 값

$$H_0: \theta \geq \theta_0 (\theta = \theta_0) \quad \text{vs} \quad H_1: \theta < \theta_0 \quad p \text{ 값} = P(W \leq w | \theta = \theta_0)$$

$$H_0: \theta \leq \theta_0 (\theta = \theta_0) \quad \text{vs} \quad H_1: \theta > \theta_0 \quad p \text{ 값} = P(W \geq w | \theta = \theta_0)$$

(예) $Y \sim N(\mu, 4)$, $H_0: \mu \leq 5$ vs $H_1: \mu > 5$

$Y=7.5$ 가 관측값이면

$$p \text{ 값} = P(Y \geq 7.5 | \mu = 5) = P(Z \geq 1.25) = 1 - \Phi(1.25) = 0.1056$$

(m) 양쪽검증의 p 값

검증통계량은 W 이고 w 는 표본으로부터 계산한 W 의 값

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_1: \theta \neq \theta_0 \quad p \text{ 값} = P(|W - \eta_0| \geq |w - \eta_0| | \theta = \theta_0)$$

여기서 $\eta_0 = h(\theta_0)$ 이고 W 는 $\eta = h(\theta)$ 의 추정량

(예) $Y \sim N(\mu, 4)$, $H_0: \mu = 5$ vs $H_1: \mu \neq 5$

관측값 $Y=7.5$, Y 는 μ 의 추정량이므로

$$p \text{ 값} = P(|Y - 5| \geq |7.5 - 5| | \mu = 5) = P(|Z| \geq 1.25)$$

$$= 2P(Z \geq 1.25) = 2(0.1056) = 0.2112$$

7. 엔트로피(entropy)

(1) 이산형 확률변수 X 의 확률분포가 $P(X = x_i) = p_i, i = 1, \dots, n$ 일 때

$$EN(X) = - \sum_{i=1}^n p_i \log p_i$$

를 확률변수 X 의 엔트로피라고 한다.

불확실성이 0 인 경우

$$p_k = 1, p_i = 0, i \neq k \text{ 이면 } EN(X) = 0$$

불확실성이 최대인 경우

$$p_i = \frac{1}{n}, i = 1, \dots, n \text{ 이면 } EN(X) = \log n$$

(2) 연속형 확률변수 X 의 확률밀도함수 $f(x)$

$$\text{Shannon 엔트로피} = - \int f(x) \log f(x) dx$$

(3) 두 확률분포의 비교

$$\text{이산형 : } D(p, q) = \sum p_i \log \left(\frac{p_i}{q_i} \right)$$

$$\text{연속형 : } D(f, g) = \int f(x) \log \left(\frac{f(x)}{g(x)} \right) dx$$