5장. 점(點)추정론

$x_1, \cdots, x_n$에 대하여 확률모형 $f_X(x\,;\theta)$를 가정하였을 때 관측 표본이 어느 파라미터에서 나왔는지를 알고자 하는 것이 θ에 대한 점추정(point estimation)입니다. 문제는 유한 크기의 표본으로는 점찍듯이 맞춘다는 것이 불가능하다는 것이지요. 따라서 점추정량의 확률적 행태로써 추정량 $\hat{\theta}$에 대한 성적을 매길 수밖에 없습니다. 그래서 이 장에서는 어떤 확률적 행태를 볼 것인가에 대하여 논의하게 됩니다.

첫째로, 가장 상식적인 평가기준인 추정에서의 비편향성(unbiasedness)을 봅니다. $\hat{\theta}$은 평균적으로 θ여야 한다는 것이지요. 둘째로는, 추정량이 안정적이여야 한다는 것입니다. 즉 추정량의 분산이 작아야 한다는 것, 최소분산성(minimum variance)입니다. 비편향성과 최소분산성이라는 두 기준을 결합하여 MSE(최소제곱오차)라는 것도 생각할 수 있습니다.

대표적인 점추정론 두 가지를 이 장에서 다루겠습니다. 하나는 최소분산비편향추정(MVUE, minimum variance unbiased estimation)의 이론으로서, 비편향추정량 가운데 분산이 최소인 추정량을 찾는 방법을 학습하게 될 것입니다. 라오(Rao)-블랙웰, 레만-쉐페, 크래머-라오(Rao) 등 여러 가이드가 여러분을 안내하게 될 것입니다.

최대가능도추정(MLE, maximum likelihood estimation)의 방법에 대하여는 이미 4장에서 소개하였습니다. 이 장에서는 MLE에 의한 추정량의 제 성질을 유도하게 됩니다. MLE는 점근적으로 (표본크기가 커짐에 따라) 정규성, 비편향성, 최소분산 등 여러 좋은 성질을 갖게 됨을 알게 될 것입니다. 큰 표본에서 이런 성질을 갖는다는 것은 작은 표본에서도 이에 근사한 성질을 가질 가능성이 있기 때문에 고무적이지요. 그러나 실제로 꼭 그런 것은 아니기 때문에 수학적 방법이나 몬테칼로 방법 등으로 소표본적 성질을 알아보는 것이 필요합니다.

차례 :

5.1 비편향성(Unbiasedness)

임의표본 자료 $x_1, \cdots, x_n$에 대하여 확률모형 $f_X(x;\theta)$를 가정하였다고 합시다. 우리가 일차적으로 알고 싶은 것은 당연히 파라미터 θ입니다. 이것은 θ가 특히 실질적으로 의미 있는 특성치인 경우에 더욱 그렇지요. 예를 들어 자료가 어떤 전자부품의 수명이라고 합시다. 그리고 이에 대한 확률모형이 Exponential(θ)라고 합시다. 이때 θ는 모분포의 평균입니다. 표본자료로부터 모평균 θ를 1개 값으로 꼭 집어 말해야 한다면 어떻게 해야 할까요? 이 점추정(點推定, point estimation) 문제는 이미 앞장에서 다루어진바 있습니다만, 그것도 한 방법일 뿐 다른 방법도 얼마든지 있을 수 있습니다. 예컨대 mle인 $\bar{x}$로 추정할 수도 있겠고 중위수(median) $\tilde{x}$로 추정할 수도 있을 것입니다. 아니면 연구자가 생각하고 있던 μ라는 특정 상수와 $\bar{x}$를 반씩 섞어서 (즉, $0.5\mu + 0.5\bar{x}$로) θ를 추정할 수도 있겠습니다.

아뭏든 $T(x_1, \cdots, x_n)$으로 θ를 점추정한다고 합시다. 추정값이 θ에 얼마만큼 근접했는가를 볼 수 있는 경우도 있겠습니다만, 1개 결과만 가지고 추정방법의 좋고 나쁨을 논할 수는 없습니다. 미아리 점쟁이가 특정 대통령 선거에서 A 후보의 득표율을 쪽집게 같이 맞추었다고 해서 그의 예지력이 훌륭하다고 인정할 수는 없는 것과 같은 이치입니다. 여러 번에 걸쳐 거듭 그의 예측이 정확하다면야 그의 능력을 인정하지 않을 수 없겠지요. 그러므로 $T(x_1, \cdots, x_n)$ 값 대신, $T(X_1, \cdots, X_n)$의 확률적 행태(behavior)가 추정방법(추정량) 평가의 토대가 됩니다.

우리가 $T = T(X_1, \cdots, X_n)$에 대하여 일차적으로 확인하고 싶은 것은 T의 기대값이 θ가 되는가 하는 것입니다. T가 때에 따라 θ를 넘어가기도 하고 때에 따라 θ에 못 미치기도 하겠지만, 평균적으로는 넘치고 모자람이 없어야 할 것입니다. 상식적인 요구입니다.

비편향성(非偏向性, unbiasedness) : 정의

추정량 T의 기대값이 θ이면, 즉

$$E(T;\theta) = \theta$$

일 때, T를 θ의 비편향추정량(unbiased estimator)이라고 합니다. '비편향'이 아닌 것은 모두 편향추정량(biased estimator)입니다.

※ 'bias'를 흔히 편의(偏倚)라고 해왔습니다만 어려운 한자이기 때문에 많은 사람들이 잘못 읽기도 하였습니다. 때문에 이 책에서는 편향(偏向)이라고만 하겠습니다.

예를 들어 $X_1, \cdots, X_n$이 Exponential(θ)로부터의 임의표본인 경우,

$$E(\overline{X};\theta) = \frac{1}{n}\{E(X_1;\theta) + \cdots + E(X_n;\theta)\} = \theta$$

입니다. 따라서 $T = \overline{X}$는 θ에 대한 비편향추정량입니다. 비편향성의 관점에서 $\overline{X}$는 괜찮은 추정량인 것이죠.

이 번에는 $X_1, \cdots, X_n$이 Uniform(θ)로부터의 임의표본이라고 합시다. 이 때, θ를 $T = \max(X_1, \cdots, X_n)$으로 추정한다고 합시다.

$$E(T;\theta) = \int_0^\theta t \cdot n\,(t/\theta)^{n-1} dt/\theta = \frac{n}{n+1}\theta < \theta$$

입니다. 따라서 T는 θ에 대한 편향추정량입니다. 사실, T는 θ에 평균적으로 못 미칠 뿐만 아니라 항상 θ에 못 미칩니다 (왜?). 따라서 비편향성의 관점에서 T는 θ의 추정량으로서 문제가 있습니다. 그러나 표본크기가 커짐에 따라 $E(T;\theta)$는 θ로 수렴합니다. 즉,

$$\lim_{n\to\infty} E(T;\theta) = \lim_{n\to\infty} \frac{n}{n+1}\theta = \theta$$

입니다. 그러므로 아주 큰 표본에서는 편향이 거의 없습니다.

점근적 비편향성(漸近的 非偏向性, asymptotic unbiasedness) : 정의

추정량 T의 기대값이 표본크기 n이 무한대로 커짐에 따라 θ에 수렴하면, 즉

$$\lim_{n\to\infty} E(T;\theta) = \theta$$

일 때, T를 θ의 점근적 비편향 추정량(asymptotically unbiased estimator)이라고 합니다.

연습으로, $X_1, \cdots, X_n$이 Exponential(θ)로부터의 임의표본인 경우,

$$\tilde{X} = \text{median}(X_1, \cdots, X_n)$$

이 θ에 대한 비편향추정량인가를 검사하여 봅시다. 중위수(median)은 여러 경우에서 평균에 대한 대안이기 때문에 평균 $\overline{X}$ 대신 중위수 $\tilde{X}$로 θ를 추정할 수 있겠는가를 생각해보자는 것입니다. $T = \tilde{X}$로 놓고 T의 확률분포를 구해봅시다. 설명을 간단히 하기 위하여 $n = 2m+1$(홀수)인 경우를 다루겠습니다. 이 경우에서는 중위수 $t = \tilde{x}$가 $m+1$번째 자료값이므로

$$f_T(t;\theta) = \frac{(2m+1)!}{m!\,m!} \cdot (1-e^{-t/\theta})^m \cdot (e^{-t/\theta})^m \cdot \frac{1}{\theta} e^{-t/\theta}, \quad t \geq 0$$

입니다 (왜?). T의 기대값은

$$E(T;\theta) = \int_0^\infty t \cdot f_T(t;\theta)\, dt$$

$$= \int_0^\infty t \cdot \frac{(2m+1)!}{m!\, m!} \cdot (1-e^{-t/\theta})^m \cdot (e^{-t/\theta})^m \cdot \frac{1}{\theta} e^{-t/\theta}\, dt$$

$$= \theta \int_0^\infty s \cdot \frac{(2m+1)!}{m!\, m!} \cdot (1-e^{-s})^m \cdot (e^{-s})^m \cdot e^{-s}\, ds$$

$$= \theta \int_0^\infty s \cdot f_T(s;1)\, ds$$

가 됩니다. 따라서, $\theta = 1$인 경우 T의 기대값인

$$I(m) = \int_0^\infty t \cdot f_T(t;1)\, dt$$

를 계산할 필요가 있습니다.

한 예로서 $n = 9$, 즉 $m = 4$인 경우에 정적분 $I(4)$를 수치적으로 계산하여 봅시다 (이 정적분은 해석적으로도 계산 가능할 것으로 예상됩니다. 한 번 해보세요). 수치적분 (numerical integration)엔 여러 방법이 있습니다만 (최영훈 · 이승천 (1995) 등 참조) 그 중 한 방법을 쓰도록 하겠습니다.

심프슨의 1/3 규칙(Simpson's 1/3 Rule) : $\int_a^b g(x)\, dx$를 수치적으로 구하기

1) 적분구간(a,b)를 N 등분합니다. 그러면 i번째 소구간을

$$(a_i, b_i),\ i = 1, \cdots, N;\ a_i = a + \frac{b-a}{N}(i-1),\ b_i = a + \frac{b-a}{N} i$$

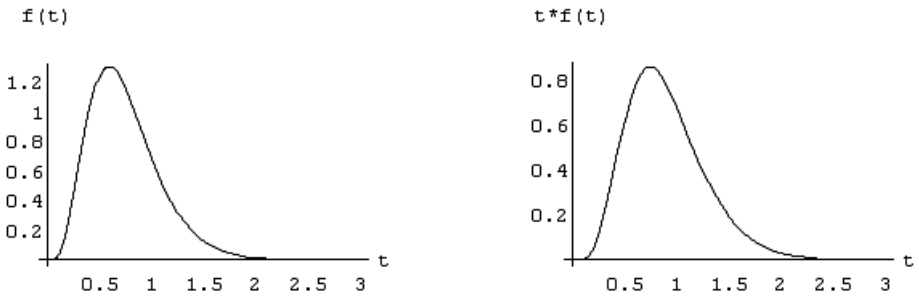
로 표현할 수 있습니다.

2) 소구간(a_i, b_i)에서의 적분 $\int_{a_i}^{b_i} g(x)\, dx$를 다음과 같이 근사시킵니다.

$$\int_{a_i}^{b_i} g(t)\, dt \simeq \frac{b_i - a_i}{3} \left[\frac{1}{2} g(a_i) + 2 \cdot g\left(\frac{1}{2}(a_i + b_i) \right) + \frac{1}{2} g(b_i) \right].$$

3) 따라서

$$\int_a^b g(x)\, dx \simeq \sum_{i=1}^{N} \frac{b_i - a_i}{6} \left[g(a_i) + 4 \cdot g\left(\frac{1}{2}(a_i + b_i) \right) + g(b_i) \right]. \qquad \blacksquare$$



<그림 1> $\theta = 1$인 경우 중위수 T의 확률밀도 $f_T(t;1)$와 $t \cdot f_T(t;1)$의 플롯

<표 1> 심프슨의 1/3 규칙에 의한 수치적분

```
/* Numerical Integration by Simpson's 1/3 Rule */
/*    g(x) = x*630*((1-exp(-x))**4)*exp(-5*x)  */
/* simpson.iml                                 */

proc iml;
    a = 0;  b = 3;
    N = 1000;

    start g;
      g1 = x1*630*((1-exp(-x1))**4)*exp(-5*x1);
      g2 = x2*630*((1-exp(-x2))**4)*exp(-5*x2);
      g3 = x3*630*((1-exp(-x3))**4)*exp(-5*x3);
    finish;

    sum = 0;
    do i=1 to N;
        x1 = (b-a)/N*(i-1);
        x3 = (b-a)/N*i;
        x2 = (x1+x3)/2;
        run g;
        sum = sum + (b-a)/N*(g1+4*g2+g3)/6;
    end;

    print sum[format=10.4];
quit;
```

<그림 1>을 보십시오. 오른 쪽 그래프가 적분해야 할 함수입니다. $a=3$으로, $b=\infty$ 대신 $b=3$으로 하겠습니다. 구간 수를 $N=1{,}000$으로 잡고서 <표 1>의 프로그램을 돌린 결과 $I(4)=0.7455$를 얻었습니다. 따라서

$$E(T;\theta)=\int_0^{\infty} t\cdot f_T(t;\theta)\,dt=0.7455\,\theta$$

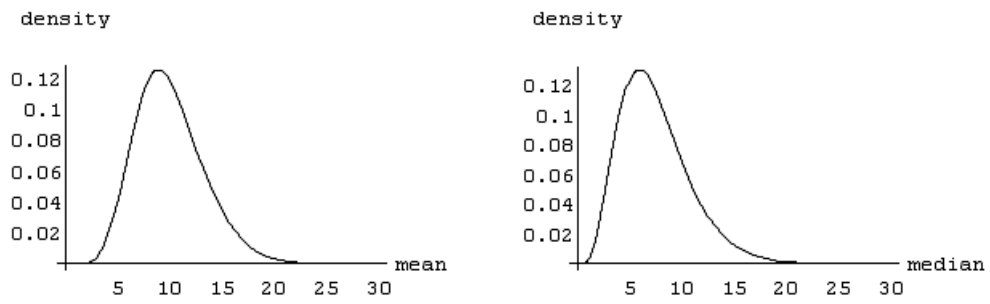
입니다 (이런 수치적분에 대한 대안은 몬테칼로 적분입니다. 연습문제 5.3). 그러므로 중위수 $T(=\tilde{X})$는 θ에 대한 편향추정량입니다. 짐작하기 쉽지 않은 결과이지요.

그렇다면 지수분포 표본의 경우에 표본크기 n이 커지면 T가 어떻게 될까요? 쉽게 직관적으로 생각해봅시다. n이 무한히 커지면 표본자체가 모집단, 즉 모분포를 형성할 것입니다. 표본이 모분포가 됨에 따라 표본 중위수 T는 모분포의 중위수 ξ가 될 것입니다. 그것은

$$F_X(\xi)=\exp\left(-\frac{\xi}{\theta}\right)=\frac{1}{2}\quad\Rightarrow\quad \xi=\theta\log_e 2=0.6931\,\theta$$

입니다. 따라서 점근적으로도 T가 θ에 대한 비편향성을 가질 것으로 기대하기는 어렵습니다. (그러나 $0.6931^{-1}\tilde{X}$는 점근적으로 비편향적일 것입니다).

이 절의 핵심은 추정량 T의 성질을 알려면 그것의 확률분포를 조사해야 할 필요가 있다는 것입니다. 특히 T의 기대값이 θ가 되는지, 즉 비편향성 여부가 관심입니다. 일반적으로, 통계량의 확률분포를 표집분포(標集分布, sampling distribution)라고 합니다. <그림 2>에서 왼쪽 그림은 $\theta=10$이고 $n=9$인 지수표본에서 $\overline{X}$의 표집분포이고 오른쪽 그림은 중위수 $\tilde{X}$의 표집분포입니다. 중위수의 기대값이 $\theta=10$에 못 미침을 볼 수 있습니다.



<그림 2> $\theta=10$인 경우 $\overline{X}$의 표집분포와 $\tilde{X}$의 표집분포

5.2 변동성(Variation)

파라미터 θ를 위한 추정량에 대하여, 비편향성 뿐만 아니라 그것의 변동성도 고려해야 합니다. 이것은 화살 쏘기의 비유로 생각하면 당연합니다. 궁사(窮士)는 햇빛 탓, 바람 탓, 기분 탓하지 말고 화살을 과녁 중심에 집중적으로 맞추어야 할 것입니다. 그렇기 때문에 흐트러짐을 나타내는 $Var(T;\theta)$를 조사할 필요가 있습니다.

예를 들어, $X_1, \cdots, X_n$이 Uniform$(0,\theta)$로부터의 iid 확률변수라고 합시다. 이 때 θ에 대한 추정량으로

$$T_1 = 2\overline{X} \quad , \quad T_2 = \frac{n+1}{n} X_{(n)}$$

를 생각한다고 합시다 (여기서 $X_{(n)} = \max\{X_1, \cdots, X_n\}$). 어느 것이 더 좋을까요?

X_1의 평균과 분산이 $\theta/2$와 $\theta^2/12$이기 때문에

$$E(T_1;\theta) = \theta, \quad Var(T_1;\theta) = \frac{\theta^2}{3n}$$

입니다. 이제 T_2의 평균과 분산을 구해보기로 합시다. $X_{(n)}$의 확률밀도는

$$f_{X_{(n)}}(x) = \frac{n}{\theta}\left(\frac{x}{\theta}\right)^{n-1}, \quad 0 \leqq x \leqq \theta$$

입니다. $E(X_{(n)};\theta) = \frac{n}{n+1}\theta$, $Var(X_{(n)};\theta) = \frac{n}{(n+1)^2(n+2)}\theta^2$이므로

$$E(T_2;\theta) = \theta, \quad Var(T_2;\theta) = \frac{\theta^2}{n(n+2)}$$

가 됩니다. 따라서 T_1과 T_2가 모두 비편향추정량입니다. 그러나, $n \geq 2$인 경우엔

$$Var(T_1;\theta) \; > \; Var(T_2;\theta)$$

입니다. 그러므로 변동성의 관점에서 T_2가 T_1에 비해 월등 나은 추정량입니다. 여기서 $Var(T_1;\theta)$ 대 $Var(T_2;\theta)$의 비를 상대적 효율성으로 정의합니다.

T_1에 비교한 T_2의 상대적 효율성(relative efficiency) : 정의

$$RE(T_2, T_1;\theta) = \frac{Var(T_1;\theta)}{Var(T_2;\theta)}. \qquad \blacksquare$$

앞의 사례에선

$$RE(T_2, T_1;\theta) = \frac{\theta^2/(3n)}{\theta^2/(n(n+2))} = \frac{n+2}{3} > 1 \quad \text{for } n \geq 2$$

입니다. 수치예로 $n = 10$인 경우엔 RE가 4입니다. 이것은 $n = 10$인 경우, T_2의 분산이 T_1의 분산의 1/4에 불과한 것을 뜻합니다.

다른 한 예로, $X_1, \cdots, X_n$이 오염정규분포(contaminated normal distribution)로부터의 iid 확률변수라고 합시다. 즉,

$$f_X(x;\theta,\sigma^2) = 0.9 \cdot \frac{1}{\sigma}\,\phi\left(\frac{x-\theta}{\sigma}\right) + 0.1 \cdot \frac{1}{k\sigma}\,\phi\left(\frac{x-\theta}{k\sigma}\right), \quad k \geq 1$$

이라고 합시다. 여기서

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \quad -\infty < z < \infty$$

입니다. 그러므로 이 오염분포는 두 정규분포 $\mathrm{N}(\theta,\sigma^2)$과 $\mathrm{N}(\theta,k^2\sigma^2)$을 0.9와 0.1의 비율로 혼합한 확률분포입니다. 여기서 $k \geq 1$이므로, $k=1$이 아닌 한, 좋은 질의 동질적인 데이터가 0.9의 비율로, 나쁜 질의 이질적인 데이터가 0.1의 비율로 자료가 섞이는 상황을 모형화한 것이라고 할 수 있지요. <그림 3>을 보십시오.

이 오염분포의 중심 θ에 대하여 두 추정량

$$T_1 = \overline{X} = \mathrm{mean}(X_1, \cdots, X_n) \quad \text{대} \quad T_2 = \widetilde{X} = \mathrm{median}(X_1, \cdots, X_n)$$

을 생각하기로 합시다. 어느 것이 더 나을까요?

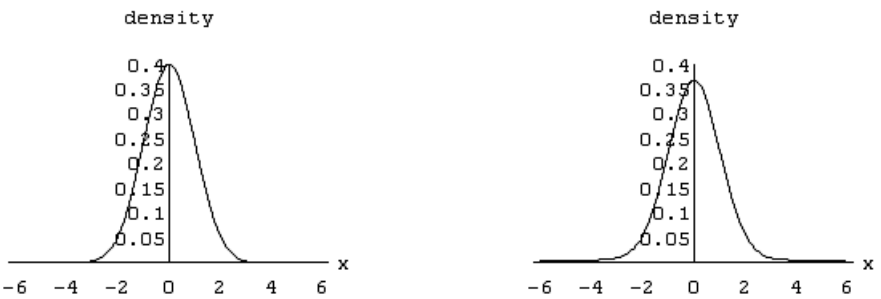
X_1의 평균이 $E(X_1;\theta,\sigma^2) = \theta$, 분산이

$$Var(X_1;\theta,\sigma^2) = 0.9 \cdot \sigma^2 + 0.1 \cdot k^2\sigma^2 = (0.9 + 0.1\,k^2)\sigma^2$$

이기 때문에

$$E(T_1;\theta) = \theta, \quad Var(T_1;\theta) = \frac{\sigma^2}{n}(0.9 + 0.1\,k^2)$$

입니다. 이제 T_2의 평균과 분산을 구해보기로 합시다. $\widetilde{X}$의 확률밀도함수가 θ를 중심으로 대칭적이라는 것은 명확합니다. 그러므로



<그림 3> 오염 정규분포($\theta=0, \sigma^2=1$) : $k=1$ (왼쪽)과 $k=5$(오른쪽)

$$E(T_2;\theta) = \theta$$

입니다. 그러나 $\tilde{X}$의 분산을 해석적으로 구하기는 어려워 보입니다 (근사적으로는

$$Var(T_2;\theta) \sim \frac{\sigma^2}{n} \cdot \frac{\pi/2}{(0.9+0.1/k)^2}$$

이라는 결과가 있습니다. Rice (1995)의 Mathematical Statistics and Data Analysis 2nd Edition, 376쪽 참조). 따라서 T_1과 T_2의 변동성을 몬테칼로 모의시행을 통해 비교해 보기로 하겠습니다.

몬테칼로 연구에서 $\theta = 10$, $\sigma^2 = 1$로 하겠습니다 (잘 따져보면 이것은 일반성을 저해하지 않는 제약입니다). 그리고 $n(=40)$개의 표본중에서 $0.9n(=36)$개의 표본은 정규분포 $N(\theta, 1)$로부터, 나머지 $0.1n(=4)$개의 표본은 정규분포 $N(\theta, k^2)$로부터 발생시켜 표본자료를 만들고 표본평균 T_1과 중위수 T_2의 값을 산출합니다. 그리고 이런 과정을 총 $N(=1{,}000)$번 반복하여 T_1과 T_2의 분산을 추정하는 몬테칼로 모의시행을 해보자는 것입니다. 이 때, k는 $1,2,3,4,5$에 대하여, n은 $10,20,40$에 대하여 해보도록 하겠습니다. <표 2>의 프로그램을 보십시오. 몬테칼로 연구의 결과를 정리해보면 다음과 같습니다.

n = 10	분 산				
	k = 1	k = 2	k = 3	k = 4	k = 5
평 균 T_1	0.1045	0.1312	0.1789	0.2725	0.3443
중위수 T_2	0.1400	0.1526	0.1626	0.1627	0.1641

n = 20	분 산				
	k = 1	k = 2	k = 3	k = 4	k = 5
평 균 T_1	0.0497	0.0655	0.0896	0.1162	0.1687
중위수 T_2	0.0709	0.0845	0.0803	0.0838	0.0842

n = 40	분 산				
	k = 1	k = 2	k = 3	k = 4	k = 5
평 균 T_1	0.0255	0.0314	0.0443	0.0623	0.0899
중위수 T_2	0.0362	0.0431	0.0420	0.0441	0.0457

표본크기에 별 관계 없이 $k=3$을 경계로 그 미만에선 T_1의 분산(변동성)이 T_2의 분산보다 더 작고, 그 이상에선 반대라고 하겠습니다. 다시 말하여 $k \geq 3$에서는 (더욱 질의 이질적인 나쁜 자료가 섞이는 경우에는) 평균 T_1보다 중위수 T_2가 더 효율적입니다.

<표 2> 오염정규분포에서 두 통계량 T_1과 T_2의 비교

```
/* Sampling from Contaminated Normal Distributions */
/* contam2.iml                                     */

proc iml;
    theta = 10;  sigma = 1;  n = 40;  n1 = n*0.9;
    Nrepeat = 1000;

    start stats;              /* Computing Median and Mean */
       r = rank(x);  x1 = j(n,1,0);
       do i=1 to n;  x1[r[i]] = x[i];  end;
       if mod(n,2) = 0 then median = (x1[n/2] + x1[n/2+1])/2;
          else median = x1[(n+1)/2];
       mean = sum(x)/n;
    finish;

    do k = 1 to 5;
       x = j(n,1,0);  M1 = j(Nrepeat,1,0);  M2 = j(Nrepeat,1,0);

       /* Monte-Carlo Generation of Contaminated Samples */
       do repeat=1 to Nrepeat;
          do i=1 to n1;    x[i] = theta + sigma*normal(0);    end;
          do i=n1+1 to n;  x[i] = theta + k*sigma*normal(0);  end;
          run stats;
          M1[repeat] = mean;  M2[repeat] = median;
       end;
       mean1 = sum(M1)/Nrepeat;
       var1 = (ssq(M1) - Nrepeat*mean1*mean1)/(Nrepeat-1);
       mean2 = sum(M2)/Nrepeat;
       var2 = (ssq(M2) - Nrepeat*mean2*mean2)/(Nrepeat-1);
       print n k mean1[format=9.2] var1[format=9.4]
                 mean2[format=9.2] var2[format=9.4];
    end;
quit;
```

5.3 MSE(평균제곱오차) 기준

이제까지 θ를 위한 추정량 T에 대하여 비편향성과 변동성을 생각해보았습니다. 그런데 추정량 T_1이 T_2에 비하여 편향은 작은데 분산이 크다면 어떤 것이 더 나을까를 결정하기가 곤란해집니다. 따라서 편향과 분산을 통합한 기준을 고려할 필요가 있습니다.

MSE (평균제곱오차, mean squared error) 기준 : 정의와 성질

$$\mathrm{MSE}(T,\theta) = E\{(T-\theta)^2;\theta\}.$$

그런데

$$\begin{aligned} E\{(T-\theta)^2;\theta\} &= E[\{T-E(T;\theta)+E(T;\theta)-\theta\}^2;\theta] \\ &= E[\{T-E(T;\theta)\}^2] + \{E(T;\theta)-\theta\}^2 \end{aligned}$$

이므로

$$\mathrm{MSE}(T,\theta) = Var(T;\theta) + \mathrm{Bias}^2(T;\theta)$$

입니다. 다시 말하여, MSE는 분산과 제곱편향의 합입니다. 여기서

$$\mathrm{Bias}(T,\theta) = E(T;\theta)-\theta$$

는 추정량 T의 편향(偏向, bias)을 뜻합니다. ■

MSE가 분산과 제곱편향의 합이므로 작은 MSE가 되기 위해서는 제곱편향과 분산이 '균형적으로' 작아져야 합니다.

한 예로서, $x_1,\cdots,x_n$이 특정 그룹에 속한 학생들의 IQ 점수라고 합시다. 이들 자료에 대하여 정규분포 $\mathrm{N}(\theta,\sigma^2)$ 모형을 적용하기로 하고, θ에 대한 추정량으로

$$T_1 = \overline{X} \quad 과 \quad T_2 = \frac{1}{2}(100+\overline{X})$$

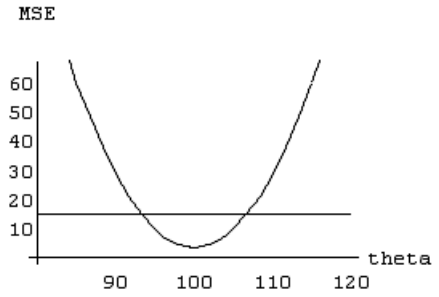
를 비교해 봅시다. T_2는 IQ의 우리나라 전체 모집단 평균이 100이라는 것을 반영한 추정량입니다. 어느 것이 더 나을까요? T_1은 비편향추정량이므로

$$\mathrm{MSE}(T_1,\theta) = Var(T_1,\theta) = \frac{\sigma^2}{n}$$

입니다. 반면, T_2는 편향되어 있습니다.

$$\mathrm{Bias}(T_2,\theta) = E(T_2;\theta)-\theta = \frac{1}{2}(100+\theta)-\theta = \frac{1}{2}(100-\theta)$$

이므로



<그림 4> T_1의 MSE (= 수평선)와 T_2의 MSE (= 포물선)

$$\mathrm{MSE}(T_2,\theta) = Var(T_2,\theta) + \mathrm{Bias}^2(T_2,\theta) = \frac{\sigma^2}{4n} + \left\{ \frac{1}{2}(100-\theta) \right\}^2$$

입니다. 분산이 $\sigma^2 = 15^2$으로 알려져 있다고 하고 $n = 15$일 때 T_1과 T_2의 MSE를 비교해봅시다. <그림 4>를 보십시오. $\theta = 100$ 근처에서는 T_2의 MSE가 T_1의 MSE에 비해 작습니다. 다시 말하여, 표본이 추출된 특정 그룹이 전체 모집단과 큰 차이가 없는 경우에는 추정량 T_2가 더 낫다는 것이죠.

다른 한 예로, $X_1, \cdots, X_n$이 정규분포 $\mathrm{N}(\theta_1, \theta_2)$로부터의 iid 확률변수라고 합시다. 그리고 $\theta_2 (= \sigma^2)$를 위한 추정량으로

$$T_1 = \sum_{i=1}^{n} (X_i - \overline{X})^2 / (n-1), \qquad T_2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 / n$$

을 생각하기로 합시다. 잘 알려져 있듯이

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 / \theta_2 \quad \sim \quad \chi^2(n-1),$$

즉 자유도가 $n-1$인 카이제곱분포를 따릅니다. 따라서 T_1은 비편향적입니다. 반면 T_2는 편향을 갖습니다. 그러나

$$Var(T_2) = \left(\frac{n-1}{n} \right)^2 Var(T_1)$$

이므로 분산은 T_2가 더 작습니다. 그러므로 MSE(평균제곱오차)로 T_1과 T_2를 비교해봅시다. 그 결과는

$$\mathrm{MSE}(T_1,\theta_1,\theta_2) = \theta_2^2 \cdot \frac{2}{n-1}, \qquad \mathrm{MSE}(T_2,\theta_1,\theta_2) = \theta_2^2 \cdot \frac{2n-1}{n^2}$$

입니다 (연습문제 5.4). 따라서 T_2의 MSE가 T_1의 MSE에 비해 항상 작습니다. 우리가 의례 T_1를 $\theta_2 (= \sigma^2)$에 대한 추정량으로 써왔기 때문에 당황스럽지 않습니까?

MSE의 기준에서 이것보다 더 좋은 것이 있는데 말입니다.

이 문제는 파라미터 θ_2가 분포의 척도(scale)와 관련이 있는 경우에는 MSE 기준이 타당한지 근본적인 의문을 시사합니다. $\theta_2 (=\sigma^2)$는 위치(location) 모수 θ_1과는 달리 기본적으로 $\theta_2 > 0$입니다. 때문에, θ_2의 과다예측과 과소예측에 동일한 페날티(penalty)를 주는 MSE 기준은 마땅하지 않습니다. 이것보다는 다음과 같은 로그변환형의 MSE 기준이 좋지 않을까요?

$$\text{MSLE}(T, \theta_1, \theta_2) = E\left\{(\log_e T - \log_e \theta_2)^2 ; \theta_1, \theta_2\right\}.$$

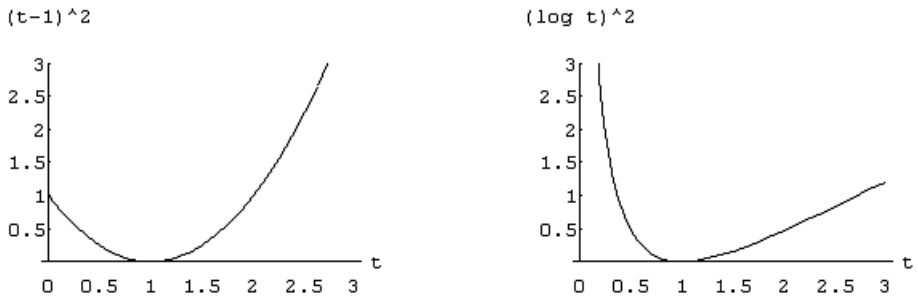
<그림 5>를 보세요. T가 $\theta_2 \cdot \chi^2(n-1)/k_n$ (k_n은 상수)의 형태인 경우에는

$$\text{MSLE}(T, \theta_1, \theta_2) = E\left[\left\{\log_e (\chi^2(n-1)/k_n)\right\}^2\right]$$

이므로 MSLE가 (θ_1, θ_2)에 무관합니다. T_1과 T_2의 MSLE를 계산해보니 다음과 같았습니다 (수치적분 결과 : 연습문제 5.5).

MSLE	n=10	n=20	n=40	n=100
T_1	0.2620	0.1139	0.05329	0.02051
T_2	0.2974	0.1220	0.05524	0.02081

이 절의 핵심 내용은 추정량을 평가하는 기준으로서 MSE(평균제곱오차)입니다. MSE는 추정론에서 대체로 인정받는 기준입니다만, 미지의 파라미터에 의존적인 경우가 있고 다른 기준이 보다 타당할 수 있다는 점에서 한계가 있습니다.



<그림 5> 제곱오차와 제곱로그오차 : $\theta_2 = 1$인 경우

5.4 MVUE(최소분산비편향추정) 이론

이 절에서는 비편향성(unbiasedness)을 점 추정량으로서의 자격요건으로 합니다. 만약 비편향추정량이 여럿 있다면 가급적 분산이 작은 것이 좋겠지요. 따라서 궁극적 목표는 비편향추정량으로서 가장 작은 분산을 갖는 것이 무엇인지 알아내는 것이겠습니다. 그것이 가능할까요? 일반적으로, 아니면 어떤 조건하에서?

당연하게도, '보다 작은' 것을 찾는 일이 '가장 작은' 것을 찾는 일보다 쉬울 것입니다. 그러므로 쉬워 보이는 문제에 먼저 도전해 봅니다.

라오-블랙웰(Rao-Blackwell) 정리 : 보다 작음을 찾아서

$X_1, \cdots, X_n$ 을 확률밀도함수 $f(x;\theta)$ 로부터의 iid 임의변수들이라고 하고, 통계량 S가 이 분포의 파라미터 θ를 위한 충분통계량이라고 합시다. T를 $\tau(\theta)$에 대한 비편향추정량이라고 할 때, 통계량

$$\phi_T(S) \equiv E(T \mid S)$$

는 다음 두 성질을 갖습니다.

1) $\phi_T(S)$ 는 T와 마찬가지로 $\tau(\theta)$에 대한 비편향추정량이다.

2) 그러나, $\phi_T(S)$ 의 분산이 T의 분산보다 작다. 즉,

$$Var(\phi_T(S);\theta) \le Var(T;\theta), \text{ 모든 } \theta \text{에 대하여.}$$

이 정리에서 핵심적 아이디어는 S에 조건화한 T의 기대값인 새 추정량 $\phi_T(S)$가 T로부터는 비편향성을 물려받고 S로부터는 충분성을 물려받아 좋은 '형질'의 합성체가 된다는 것이겠습니다. 증명은 다음과 같습니다. 1)은

$$E(\phi_T(S);\theta) = E(E(T \mid S);\theta) = E(T;\theta) = \tau(\theta)$$

로부터 나옵니다. 여기서 $E(E(T \mid S);\theta) = E(T;\theta)$인 이유는

$$\int \left(\int t \, f_{T|S}(t \mid s) \, dt \right) f_S(s;\theta) \, ds = \int\int t \, f_{T,S}(t,s;\theta) \, dt \, ds$$

이기 때문입니다 (이산형분포의 경우엔 $\int$ 을 $\sum$ 로 대체). 2)는 항등식

$$Var(T;\theta) = E\{Var(T \mid S);\theta\} + Var\{E(T \mid S);\theta\}$$

로부터 (연습문제 5.7)

$$Var(T;\theta) \ge Var\{E(T \mid S);\theta\} = Var(\phi_T(S);\theta)$$

가 유도되기 때문입니다 (등호는 T가 S의 함수인 경우 성립함).

예를 들어 $X_1, \cdots, X_n$ 을 포아송분포 Poisson(θ) 로부터의 확률표본이라고 합시다. 이 때 $\tau(\theta) = P\{X_1 = 0 ; \theta\} = e^{-\theta}$ 에 대한 비편향추정량으로

$$T = \sum_{i=1}^{n} 1(X_i = 0)/n,$$

즉 0인 관측치의 표본비율을 생각할 수 있겠습니다. 그런데 포아송 표본에서는

$$S = \sum_{i=1}^{n} X_i$$

가 파라미터 θ에 대한 충분통계량입니다. 그러므로, 라오-블랙웰 정리에 따라

$$\phi_T(S) = E(T \mid S)$$

가 T를 개선시킨 비편향추정량입니다. $S = s$ 가 주어졌을 때 X_1 의 조건부확률분포가 이항분포 B($s, 1/n$) 이므로 (왜?),

$$E(T \mid S = s) = \frac{1}{n}\sum_{i=1}^{n} P\{X_1 = 0 \mid S = s\} = \left(1 - \frac{1}{n}\right)^s$$

가 됩니다. 따라서

$$\phi_T(S) = \left(1 - \frac{1}{n}\right)^S$$

가 $e^{-\theta}$에 대한 개선된 비편향추정량입니다. 왜 $\phi_T(S)$ 가 합당한 추정량인지 생각해봅시다. 대수의 법칙에 의하여 n이 무한히 커짐에 따라 $S/n = \overline{X}$ 가 θ로 수렴하므로

$$\phi_T(S) = \left(1 - \frac{1}{n}\right)^{n\overline{X}} = \left\{\left(1 - \frac{1}{n}\right)^{n}\right\}^{\overline{X}} \rightarrow (e^{-1})^{\theta} \ (= e^{-\theta})$$

가 됩니다 (참고 : $\lim_{h \to 0} (1 \pm h)^{\frac{1}{h}} = e^{\pm 1}$).

이렇게 추정량의 확률적 극한값이 그것이 추정하려된 목표가 되는 경우 그러한 추정량을 일치추정량(consistent estimator)라고 합니다. 즉 $\phi_T(S)$ 는 $e^{-\theta}$에 대한 일치추정량입니다.

이제 가장 작은 분산을 갖는 비편향추정량을 찾아 나설 차례입니다. T_1과 T_2 를 $\tau(\theta)$에 대한 비편향추정량이라고 합시다. 이 때 일반성을 잃지 않고

$$Var(T_1 ; \theta) \leq Var(T_2 ; \theta)$$

로 가정하겠습니다. 그리고 파라미터 θ를 위한 충분통계량 S를 활용하여 이들을 개선시킨 비편향추정량을 $\phi_{T_1}(S)$ 와 $\phi_{T_2}(S)$ 라고 합시다. 즉,

$$\phi_{T_1}(S) = E(T_1 \mid S), \quad \phi_{T_2}(S) = E(T_2 \mid S)$$

라고 합시다. 라오-블랙웰 정리에 따라 $\phi_{T_1}(S)$ 와 $\phi_{T_2}(S)$ 는 모두 $\tau(\theta)$ 에 대한 비편향추정량이고 $\phi_{T_1}(S)$ 은 T_1 에 비교하여, 그리고 $\phi_{T_2}(S)$ 는 T_2 에 비교하여 작은 분산을 갖습니다. 혹시 $\phi_{T_1}(S)$ 이 $\phi_{T_2}(S)$ 보다 상대적으로 작은 분산을 갖는 것이 아닐까요?

예컨대 앞의 포아송 문제에서

$$T_1 = \sum_{i=1}^{n} 1(X_i = 0)/n, \quad T_2 = 1(X_1 = 0)$$

이라고 합시다. 이 경우 당연하게 T_1 의 분산이 T_2 의 분산보다 작습니다. 그러나 어렵지 않게

$$\phi_{T_1}(S) = \phi_{T_2}(S) \quad (= (1-n^{-1})^S)$$

임을 알 수 있습니다. 그러므로 일반적으로 $\phi_{T_1}(S)$ 가 $\phi_{T_2}(S)$ 보다 상대적으로 작은 분산을 갖는 것은 아닙니다. 그렇다면, 혹시 임의의 두 비편향추정량 T_1 과 T_2 를 충분통계량 S 로 개선시키면 그 결과가 항상 같아지는 것일까요? 이에 대한 답은, 어떤 분포의 경우에서는, 그렇다는 것입니다.

<u>완비통계량</u>(complete statistic) S : 정의

$$E\{g(S);\theta\} = 0 \quad \text{forall } \theta \quad \Rightarrow \quad g(S) = 0 \quad (\text{with probability 1})$$

인 성질이 있으면 S를 완비통계량이라고 합니다.

예를 들어, Poisson(θ) 임의표본 $X_1, \cdots, X_n$ 에서 $S = \sum_{i=1}^{n} X_i$ 가 완비통계량인지를 조사하여 봅시다. $S \sim$ Poisson($n\theta$) 이므로

$$E[g(S);\theta] = \sum_{s=0}^{\infty} g(s)\, e^{-n\theta} \frac{(n\theta)^s}{s!} = 0 \quad \text{for all } \theta$$

$$\Rightarrow \quad \sum_{s=0}^{\infty} g(s) \frac{(n\theta)^s}{s!} = 0 \quad \text{for all } \theta$$

$$\Rightarrow \quad \theta^s \text{ 항의 계수 } \quad g(s)\frac{n^s}{s!} = 0 \quad \text{for } s = 0,1,2,3, \cdots$$

$$\Rightarrow \quad g(s) = 0 \quad \text{for } s = 0,1,2,3, \cdots$$

가 됩니다. 따라서 S는 완비통계량입니다.

이번에는, Uniform(θ) 임의표본 $X_1, \cdots, X_n$ 에서 $S = X_{(n)} = \max(X_1, \cdots, X_n)$ 이 완비통계량인지를 조사하여 봅시다. S 의 확률밀도함수가

$$f_S(s;\theta) = \frac{n}{\theta}\left(\frac{s}{\theta}\right)^{n-1},\ 0 \le s \le \theta$$

이므로

$$\begin{aligned} E\{g(S);\theta\} &= \int_0^\theta g(s)\ \frac{n}{\theta}\left(\frac{s}{\theta}\right)^{n-1} ds = 0, \quad \text{for all } \theta > 0 \\ &\Rightarrow \int_0^\theta g(s)\ s^{n-1}\, ds = 0, \quad \text{for all } \theta > 0 \\ &\Rightarrow g(\theta)\theta^{n-1} = 0, \quad \text{for all } \theta > 0 \quad (\because \text{양변을 } \theta\text{로 미분함으로써}) \\ &\Rightarrow g(\theta) = 0, \quad \text{for all } \theta > 0 \end{aligned}$$

입니다. 따라서 S는 완비통계량입니다.

레만-쉐페(Lehmann-Scheffé) 정리 : 가장 작음과 유일성

충분통계량 S가 또한 완비적이라고 합시다. 그러면 $\phi(S)$는 기대값

$$E\{\phi(S);\theta\} = \tau(\theta)$$

에 대하여 유일한 최소분산비편향추정량(MVUE, minimum variance unbiased estimator)입니다. 즉 $\phi(S)$ 혼자만이 비편향추정량 중에서 최소분산을 갖습니다.

이 정리의 증명은 통계량 S의 완비충분성으로부터 유도됩니다. T를 $\tau(\theta)$에 대한 임의의 비편향추정량이라고 하면, 라오-블랙웰 정리에 의하여

$$Var(T;\theta) \ge Var(\psi_T(S);\theta), \quad \text{여기서 } \psi_T(S) \equiv E(T \mid S)$$

인데,

$$E\{\phi(S) - \psi_T(S);\theta\} = \tau(\theta) - \tau(\theta) = 0, \text{ for all } \theta.$$

그런데 S가 완비통계량이므로

$$\phi(S) - \psi_T(S) = 0 \text{ with probability 1.}$$

즉 $\phi(S) \equiv \psi_T(S)$이라는 것을 알 수 있습니다. 다시 말하여 S의 함수로서는 $\tau(\theta)$에 대하여 유일한 비편향추정량입니다. 결국,

$$Var(\phi(S);\theta) = Var(\psi_T(S);\theta) \le Var(T;\theta)$$

(이 때 등호는 $T = \psi_T(S)$, with probability 1일 때 성립)

가 됩니다. 따라서 $\phi(S)$는 $\tau(\theta)$에 대한 최소분산비편향추정량(MVUE)입니다. 만약 같은 분산을 갖는 비편향추정량이 있으면 그것은 $\phi(S)$와 같습니다(1의 확률로). 그러므로 $\phi(S)$는 유일한 MVUE입니다. ■

레만-쉐페 정리로부터, Poisson(θ) 분포로부터의 임의표본에서 $(1-n^{-1})^S$가 $P\{\mathrm{X}_1 = 0\,;\theta\} = e^{-\theta}$에 대하여 유일한 MVUE(최소분산비편향추정량)라는 사실을 알 수 있습니다. $S = \sum_{i=1}^{n} X_i$ 가 완비충분통계량이기 때문입니다.

다른 한 예로서 Uniform(θ) 임의표본에서 θ에 대한 MVUE를 구해봅시다. 이 사례에서는 $S = X_{(n)}$이 완비충분통계량입니다. 그런데 $X_{(n)}$의 기대값이

$$E(X_{(n)}\,;\theta) = \frac{n}{n+1}\,\theta$$

이므로, 레만-쉐페 정리에 의하여 $\frac{n+1}{n}\,X_{(n)}$이 θ에 대한 MVUE입니다.

또 다른 예로서, $X_1, \cdots, X_n$이 정규분포 $\mathrm{N}(\theta, \sigma^2)$로부터의 임의표본인 경우($\sigma^2$은 旣知의 定數)에서

$$\tau(\theta) = P\{\,\mathrm{X}_1 \le \mathrm{c}\,;\theta\,\} = \Phi\left(\frac{\mathrm{c}-\theta}{\sigma}\right)$$

에 대한 MVUE를 구해봅시다. 이 사례에서는 충분통계량 $\overline{X}\,(= S)$가 완비적입니다. 왜냐하면

$$E\{\,g\,(S)\,;\theta\,\} = \int_{-\infty}^{+\infty} g(s)\,\frac{1}{\sqrt{2\,\pi\,\sigma^2/n}}\,\exp\left[-\frac{(s-\theta)^2}{2\,\sigma^2/n}\right] r d\,s \;=\; 0 \quad \text{for all } \theta.$$

$$\Rightarrow \quad \int_{-\infty}^{+\infty} g(s)\,\exp\left[-\frac{s^2}{2\,\sigma^2/n}\right]\,\exp\left[\frac{s\;\theta}{\sigma^2/n}\right]\,d\,s \;=\; 0 \quad \text{for all } \theta.$$

$$\Rightarrow \quad g(s)\,\exp\left[-\frac{s^2}{2\,\sigma^2/n}\right] \;=\; 0 \quad (\;\because \text{라플라스 변환의 유일성}\;) \quad \Rightarrow \quad g(s) = 0\,.$$

그런데 $T = 1\,(X_1 \le c\,)$가 $\tau(\theta)$에 대한 비편향추정량이므로

$$\phi_T(S) = E\{T \mid S\} = P\{\,X_1 \le c \mid S\}$$

를 계산해야 합니다. 그런데

$$(X_1\,,\,S) \;\sim\; \mathrm{BN}\left(\begin{pmatrix}\theta\\ \theta\end{pmatrix}, \begin{pmatrix}\sigma^2 & \sigma^2/n\\ \sigma^2/n & \sigma^2/n\end{pmatrix}\right)$$

이므로

$$X_1 \mid S = s \;\sim\; \mathrm{N}\left(s,\, \frac{n-1}{n}\sigma^2\right)$$

가 유도됩니다 (연습문제 5.8). 따라서 MVUE는 구체적으로 다음과 같습니다.

$$\phi_T(\overline{X}) = \Phi\left(\sqrt{\frac{n}{n-1}}\;\cdot\;\frac{c-\overline{X}}{\sigma}\right).$$

이제까지의 예들에서는 충분통계량이 완비적이었지만 물론 그렇지 않은 경우들도 있습니다. 예를 들어, $X_1, \cdots, X_n$을 균일분포 Uniform$(\theta-\frac{1}{2}, \theta+\frac{1}{2})$로부터의 임의표본이라고 합시다. 그러면 $(X_{(1)}, X_{(n)})$이 θ에 대하여 결합충분합니다. 그러나

$$E(X_{(1)};\theta) = \theta - \frac{1}{2} + \frac{1}{n+1}, \quad E(X_{(n)};\theta) = \theta + \frac{1}{2} - \frac{1}{n+1}$$

이므로

$$E(X_{(n)} - X_{(1)} - 1 + \frac{2}{n+1};\theta) = 0$$

입니다. 그렇지만 꼭 $X_{(n)} - X_{(1)} = 1 - \frac{2}{n+1}$ 인 것은 아니므로 $(X_{(1)}, X_{(n)})$은 완비적이지 않습니다.

대표적으로, 충분통계량이 완비적인 경우는 지수족의 확률분포로부터의 임의표본에서입니다.

<u>지수족(exponential family)과 완비성</u> : 정리

$X_1, \cdots, X_n$이 지수족의 확률분포

$$f(x;\theta_1, \cdots, \theta_k) = \exp\left\{\sum_{j=1}^{l} c_j(\theta_1, \cdots, \theta_k) P_j(x) + d(\theta_1, \cdots, \theta_k) + Q(x)\right\},$$

for $x \in A$ (여기서 함수의 정의역 A는 $(\theta_1, \cdots, \theta_k)$와 무관)

로부터의 임의표본이면, 결합충분통계량

$$\left(\sum_{i=1}^{n} P_1(X_i), \cdots, \sum_{i=1}^{n} P_l(X_i)\right)$$

은, $k = l$ 인 경우, 완비적입니다.

증명은 라플라스 변환의 유일성에 의한 것이나 여기서는 생략합니다. ■

예를 들어, $X_1, \cdots, X_n$이 정규분포 $\mathrm{N}(\theta_1, \theta_2)$로부터의 임의표본인 경우,

$$(\overline{X}, \sum_{i=1}^{n} X_i^2), \quad \text{또는} \quad (\overline{X}, \sum_{i=1}^{n} (X_i - \overline{X})^2), \quad \text{또는} \quad (\overline{X}, S^2)$$

은 결합충분하면서 완비적입니다. 따라서 θ_1과 θ_2에 대한 MVUE는 각각

$$\overline{X} \quad \text{와} \quad S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 / (n-1)$$

이고, 분포의 상위 α 분위수 $\theta_1 + z_\alpha \theta_2^{1/2}$에 대한 MVUE는 $\overline{X} + z_\alpha c_n S$ 입니다 (여기서 c_n은 적당한 상수: 연습문제 5.9).

MVUE와는 약간 다른 관점에서 비편향추정량의 분산이 최소 얼마인가라는 것을 말해주는 부등식이 있습니다.

크래머-라오 부등식(Cramér-Rao inequality) : 비편향추정량의 분산

$X_1, \cdots, X_n$이 밀도함수 $f(\theta), \theta \in \Theta$의 확률분포로부터의 임의표본일 때, $\tau(\theta)$에 대한 비편향추정량 $T = T(X_1, \cdots, X_n)$ (즉, $E(T;\theta) = \tau(\theta)$)의 분산은, 몇 개의 정칙조건하에서

$$Var(T;\theta) \geq \frac{1}{n} \cdot \frac{\{\tau'(\theta)\}^2}{I_1(\theta)} \qquad (1)$$

입니다. 여기서

$$I_1(\theta) = E\left[\left(\frac{\partial}{\partial\theta} log\, f(X_1;\theta)\right)^2;\theta\right] \qquad (2)$$

은 단위 피셔 정보량(unit Fisher information)이고, 이 부등식을 성립하게 하는 정칙조건(正則條件; regularity conditions)은 다음 두 가지입니다.

i) $\{x \mid f(x;\theta) > 0\}$은 θ에 의존하지 않게 표현가능하다.

ii) θ에 의한 적분 $\int \cdots \int T(x_1, \cdots, x_n) \prod_{i=1}^{n} f(x_i;\theta) \cdot dx_1 \cdots dx_n$ 의 편미분은 편미분후 적분과 같다. 즉,

$$\frac{\partial}{\partial\theta} \int \cdots \int T(x_1, \cdots, x_n) \prod_{i=1}^{n} f(x_i;\theta) \cdot dx_1 \cdots dx_n$$

$$= \int \cdots \int T(x_1, \cdots, x_n) \frac{\partial}{\partial\theta} \prod_{i=1}^{n} f(x_i;\theta) \cdot dx_1 \cdots dx_n .$$

부등식 (1)의 증명은 다음과 같습니다.

$$U = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} log\, f(X_i;\theta)$$

로 정의하고 (U는 통계량이 아님), T와 U의 공분산 $Cov(T, U;\theta)$를 생각해 봅시다.

$$E(U;\theta) = \int \cdots \int \sum_{i=1}^{n} \frac{\partial}{\partial\theta} log\, f(x_i;\theta) \prod_{i=1}^{n} f(x_i;\theta)\, dx_1 \cdots dx_n$$

$$= \int \cdots \int \sum_{i=1}^{n} \left\{\frac{\partial}{\partial\theta} f(x_i;\theta) \prod_{i' \neq i} f(x_{i'};\theta)\right\} dx_1 \cdots dx_n$$

$$= \frac{\partial}{\partial\theta}\int\cdots\int\prod_{i=1}^{n} f(x_i;\theta)\ dx_1\cdots dx_n = \frac{\partial}{\partial\theta}1 = 0$$

(↑ 조건 i 에 의하여 적분영역이 θ 와 무관하므로)

이므로

$$Cov(T,U;\theta) = E[(T-\tau(\theta))\cdot U;\theta] = E[T\,U;\theta]$$

입니다. 그런데

$$\begin{aligned}
&E(T\,U;\theta)\\
&= \int\cdots\int T(x_1,\cdots,x_n)\sum_{i=1}^{n}\frac{\partial}{\partial\theta}log\,f(x_i;\theta)\prod_{i=1}^{n} f(x_i;\theta)\ dx_1\cdots dx_n\\
&= \int\cdots\int T(x_1,\cdots,x_n)\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\theta}f(x_i;\theta)\prod_{i'\neq i} f(x_{i'};\theta)\right\}dx_1\cdots dx_n\\
&= \frac{\partial}{\partial\theta}\int\cdots\int T(x_1,\cdots,x_n)\prod_{i=1}^{n} f(x_i;\theta)\ dx_1\cdots dx_n\\
&\quad (\uparrow \text{ 조건 ii 에 의하여 미분과 적분의 순서가 교환가능하므로})\\
&= \frac{\partial}{\partial\theta}\tau(\theta) = \tau'(\theta)
\end{aligned}$$

입니다. 그리고, 상관부등식

$$\frac{\{Cov(T,U;\theta)\}^2}{Var\{T;\theta\}\cdot Var\{U;\theta\}} \leq 1$$

로부터

$$Var\{T;\theta\} \geq \frac{\{Cov(T,U;\theta)\}^2}{Var\{U;\theta\}}$$

가 됩니다. 그런데,

$$\begin{aligned}
&Cov(T,U;\theta) = \tau'(\theta),\\
&Var(U;\theta) = Var\left\{\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta);\theta\right\}\\
&\qquad = \sum_{i=1}^{n} Var\left\{\frac{\partial}{\partial\theta}\log f(X_i;\theta);\theta\right\}\\
&\qquad = n\cdot Var\left\{\frac{\partial}{\partial\theta}\log f(X_1;\theta);\theta\right\} = n\cdot I_1(\theta)
\end{aligned}$$

$$\left(\because\ X_1,\cdots,X_n\text{이 독립이고}\quad E\left\{\frac{\partial}{\partial\theta}\log f(X_1;\theta);\theta\right\} = 0\right)$$

입니다. 그러므로 부등식 (1)을 얻습니다. ■

피셔 정보량의 계산 :

단위 피셔 정보량 $I_1(\theta)$는 (2) 대신

$$I_1(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1;\theta);\theta\right]$$

로도 계산가능합니다. 그 이유는 다음과 같습니다.

$$\int f(x_1;\theta)\,dx_1 = 1 \Rightarrow \frac{\partial}{\partial\theta}\int f(x_1;\theta)\,dx_1 = 0$$

$$\Rightarrow \int \frac{\partial}{\partial\theta} f(x_1;\theta)\,dx_1 = 0$$

$$\Rightarrow \int \frac{\partial}{\partial\theta}\log f(x_1;\theta)\cdot f(x_1;\theta)\,dx_1 = 0.$$

다시 양변을 θ로 편미분하면,

$$\frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta}\log f(x_1;\theta)\cdot f(x_1;\theta)\,dx_1 = 0$$

$$\Rightarrow \int \frac{\partial^2}{\partial^2\theta}\log f(x_1;\theta)\cdot f(x_1;\theta)\,dx_1 + \int\left(\frac{\partial}{\partial\theta}\log f(x_1;\theta)\right)^2 f(x_1;\theta)\,dx_1 = 0$$

이기 때문입니다.

또한, 크기 n의 iid 임의표본에서, 피셔 정보량은

$$I_n(\theta) = E\left[\left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta)\right)^2;\theta\right]$$

으로 정의됩니다. 단위 정보량과는

$$I_n(\theta) = n\cdot I_1(\theta)$$

의 관계가 있습니다 (연습문제 5.10). ■

크래머-라오 부등식(Cramér-Rao inequality)에서 등호가 성립할 조건 :

부등식 (1)에서 등호가 성립할 필요충분조건은 상관 부등식

$$\frac{\{Cov(T,U;\theta)\}^2}{Var\{T;\theta\}\cdot Var\{U;\theta\}} \le 1$$

에서 등호가 성립하는 경우입니다. 즉 $T(X_1,\cdots,X_n)-\tau(\theta)$가 U의 상수배인 경우입니다. 다시 말하여,

$$T(X_1, \cdots, X_n) - \tau(\theta) = k \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i ; \theta) \qquad (3)$$

입니다 (여기서 $k = k(\theta, n)$은 상수). 그러나, 조건 (3)이 만족되는 추정량 T가 항상 존재하는 것은 아닙니다. 다시 말하여, 이루어질 수 없는 꿈인 경우가 있다는 것입니다. 예를 곧 들어 보이겠습니다. ■

예를 들어, Poisson(θ) 임의표본 $X_1, \cdots, X_n$의 경우를 생각해봅시다.

$$f(x ; \theta) = e^{-\theta} \theta^x / x!, \quad x = 0, 1, 2, \cdots$$

이므로

$$\log f(x ; \theta) = -\theta + x \log\theta - \log x!$$

$$\Rightarrow \frac{\partial}{\partial \theta} \log f(x ; \theta) = -1 + \frac{x}{\theta}$$

$$\Rightarrow \frac{\partial^2}{\partial \theta^2} \log f(x ; \theta) = -\frac{x}{\theta^2}$$

입니다. 따라서 단위 피셔 정보량은

$$I_1(\theta) = -E\left(-\frac{X_1}{\theta^2} ; \theta \right) = \frac{1}{\theta}$$

입니다. 그러므로 크래머-라오 부등식은 이 경우

$$Var(T ; \theta) \geq \frac{\theta}{n} \cdot \{\tau'(\theta)\}^2$$

가 되겠습니다. $\tau(\theta) = \theta$인 경우엔 크래머-라오 부등식의 하한(下限, lower bound)이 곧 θ / n이 되는데 이것은 θ에 대한 비편향추정량인 $T = \overline{X}$의 분산입니다. 그러므로, 레만-쉐페 정리에 의하지 않고도 $\overline{X}$가 MVUE라는 것을 보인 것입니다. 한편 크래머-라오 부등식이 등호를 취할 조건인 식 (3)을 검토해보면

$$T - \theta = k(\theta, n) \sum_{i=1}^{n} \left(-1 + \frac{X_i}{\theta} \right) = -k(\theta, n) \frac{n}{\theta} (\overline{X} - \theta)$$

이므로 $T = \overline{X}$로하면 이 식이 만족됩니다. 따라서 $\overline{X}$의 분산이 크래머-라오 부등식의 하한이 된다는 것을 재삼 확인할 수 있습니다.

이번에는 앞의 예에서 $\tau(\theta) = e^{-\theta}$ $(= P\{X_1 = 0 ; \theta\})$에 대한 비편향추정량의 분산에 대하여 생각해보겠습니다. 크래머-라오 부등식이 등호를 취할 조건 (3)은

$$T - e^{-\theta} = -k(\theta, n) \frac{n}{\theta} (\overline{X} - \theta)$$

입니다만, T와 $k(\theta, n)$을 어떻게 잡더라도 성립하기 어렵겠다는 것을 알 수 있습니다.

한편, $\tau'(\theta) = -e^{-\theta}$ 이므로 크래머-라오 부등식은

$$Var(T;\theta) \geq \frac{\theta}{n} e^{-2\theta}$$

이 됩니다. 그러나 우리가 앞서 구한 $e^{-\theta}$에 대한 MVUE $\phi_T(S) = \left(1-\frac{1}{n}\right)^S$의 분산은, $S = n\overline{X}$ 가 Poisson$(n\theta)$를 따르므로,

$$Var\{\phi_T(S);\theta\} = (e^{\frac{\theta}{n}} - 1)\, e^{-2\theta} > \frac{\theta}{n} e^{-2\theta}, \quad \text{for } \theta > 0$$

임을 보일 수 있습니다 (연습문제 5.11). 그러므로 크래머-라오 부등식의 하한은 어떤 비편향추정량에 의하여도 달성되지 않습니다.

$X_1, \cdots, X_n$이 균일분포 Uniform$\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$로부터의 임의표본인 경우를 생각해봅시다. 크래머-라오 부등식을 적용할 수 있을까요? 그렇지 않습니다. 정칙조건 i이 충족되지 않기 때문입니다. 즉

$$\{\, x \mid f(x;\theta) > 0 \,\} = \{\, x \mid \theta - 0.5 \leq x \leq \theta + 0.5 \,\}$$

로 θ에 의존적입니다.

이제 몇 가지 분포에 대한 단위 피셔 정보량을 제시합니다 (연습문제 5.12).

- 포아송 분포 Poisson(θ) : $I_1(\theta) = \frac{1}{\theta}$,
- 베르누이 분포 Bernoulli(θ) : $I_1(\theta) = \frac{1}{\theta(1-\theta)}$,
- 지수분포 Exponential(θ) : $I_1(\theta) = \frac{1}{\theta^2}$,
- 감마분포 Gamma(α,θ) (α 既知) : $I_1(\theta) = \frac{\alpha}{\theta^2}$,
- 정규분포 N(θ,σ^2) (σ^2 既知) : $I_1(\theta) = \frac{1}{\sigma^2}$.

한편, 정규분포 N(θ_1,θ_2)와 같이 2개의 파라미터를 갖는 분포로부터의 임의표본에서의 비편향추정량에 대하여 확장된 크래머-라오 부등식은 다음과 같습니다.

크래머-라오 부등식(Cramér-Rao inequality) : 파라미터가 2개인 경우

$X_1, \cdots, X_n$이 밀도함수 $f(\theta_1,\theta_2), (\theta_1,\theta_2) \in \Theta$의 확률분포로부터의 임의표본일 때, $\tau(\theta_1,\theta_2)$에 대한 비편향추정량 $T = T(X_1, \cdots, X_n)$의 분산은, 파라미터가 1개인 경우와 비슷한 정칙조건하에서

$$Var\,(T;\theta_1,\theta_2) \;\geq\; \frac{1}{n}\nabla\tau(\theta_1,\theta_2)^t\,\{\,\mathrm{I}\,(\theta_1,\theta_2)\}^{-1}\,\nabla\tau(\theta_1,\theta_2) \qquad (4)$$

입니다. 여기서

$$\nabla\,\tau(\theta_1,\theta_2) \;=\; \begin{pmatrix} \dfrac{\partial}{\partial\theta_1}\tau(\theta_1,\theta_2) \\ \dfrac{\partial}{\partial\theta_2}\tau(\theta_1,\theta_2) \end{pmatrix} : \quad 2\times 1 \text{ 기울기 벡터,}$$

$$I(\theta_1,\theta_2) \;=\; \begin{pmatrix} I_{11}(\theta_1,\theta_2) & I_{12}(\theta_1,\theta_2) \\ I_{21}(\theta_1,\theta_2) & I_{22}(\theta_1,\theta_2) \end{pmatrix} : \quad 2\times 2 \text{ 단위 피셔 정보행렬.}$$

즉

$$\begin{aligned} I_{ij}(\theta_1,\theta_2) \;&=\; E\left[\left(\frac{\partial}{\partial\theta_i}\log f\,(X_1\,;\theta_1,\theta_2)\right)\left(\frac{\partial}{\partial\theta_j}\log f\,(X_1\,;\theta_1,\theta_2)\right);\theta_1,\theta_2\right] \\ &=\; E\left[-\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log f(X_1\,;\theta_1,\theta_2)\,;\theta_1,\theta_2\right],\; i,j=1,2\,. \end{aligned}$$

파라미터 수가 3개 이상인 경우에도 크래머-라오 부등식이 (4)은 형태로 표현됩니다. 이에 대한 증명은 생략하겠습니다. ■

예를 들어 $X_1,\cdots,X_n$ 이 $\mathrm{N}(\theta_1,\theta_2)$ 로부터의 임의표본인 경우

$$\log f(x\,;\theta_1,\theta_2) \;=\; -\frac{1}{2}\log(2\,\pi\,\theta_2) - \frac{(x-\theta_1)^2}{2\,\theta_2}$$

$$\Rightarrow \frac{\partial}{\partial\theta_1}\log f\,(x\,;\theta_1,\theta_2) \;=\; \frac{x-\theta_1}{\theta_2}\,,\quad \frac{\partial}{\partial\theta_2}\log f\,(x\,;\theta_1,\theta_2) \;=\; -\frac{1}{2\theta_2} + \frac{(x-\theta_1)^2}{2\,\theta_2^2}\,.$$

$$\Rightarrow \frac{\partial^2}{\partial\theta_1^2}\log f\,(x\,;\theta_1,\theta_2) \;=\; -\frac{1}{\theta_2}\,,\qquad \frac{\partial^2}{\partial\theta_2^2}\log f\,(x\,;\theta_1,\theta_2) \;=\; \frac{1}{2\theta_2^2} - \frac{(x-\theta_1)^2}{\theta_2^3}\,,$$

$$\frac{\partial^2}{\partial\theta_1\,\partial\theta_2}\log f\,(x\,;\theta_1,\theta_2) \;=\; -\frac{x-\theta_1}{\theta_2^2}\,.$$

$$\Rightarrow\quad I(\theta_1,\theta_2) \;=\; \begin{pmatrix} \dfrac{1}{\theta_2} & 0 \\ 0 & \dfrac{1}{2\theta_2^2} \end{pmatrix}$$

입니다. 따라서 정규분포의 분산인 $\tau(\theta_1,\theta_2)=\theta_2$ 에 대한 비편향추정량의 분산에 대한 크래머-라오 부등식은

$$Var(T;\theta_1,\theta_2) \geq \frac{1}{n}\nabla\tau(\theta_1,\theta_2)^t\{\mathrm{I}(\theta_1,\theta_2)\}^{-1}\nabla\tau(\theta_1,\theta_2)$$

$$= \frac{1}{n}(0\ \ 1)\begin{pmatrix}\frac{1}{\theta_2} & 0\\ 0 & \frac{1}{2\theta_2^2}\end{pmatrix}^{-1}\begin{pmatrix}0\\1\end{pmatrix} = \frac{2\theta_2^2}{n}$$

이 됩니다. θ_2에 대한 비편향추정량인 S^2의 분산이

$$Var(S^2;\theta_1,\theta_2) = Var\left(\theta_2\cdot\frac{\chi^2_{n-1}}{n-1}\right) = \frac{2\theta_2^2}{n-1}$$

이므로 크래머-라오 하한을 약간 초과한다는 것을 알 수 있습니다.

앞의 예에서 분포의 상위 α 분위수 $\tau(\theta_1,\theta_2)=\theta_1+z_\alpha\theta_2^{1/2}$에 대한 크래머-라오 하한은

$$\frac{1}{n}\left(1,\ z_\alpha\theta_2^{-1/2}/2\right)\begin{pmatrix}\frac{1}{\theta_2} & 0\\ 0 & \frac{1}{2\theta_2^2}\end{pmatrix}^{-1}\begin{pmatrix}1\\ z_\alpha\theta_2^{-1/2}/2\end{pmatrix} = \frac{\theta_2}{n}\left(1+\frac{z_\alpha^2}{2}\right)$$

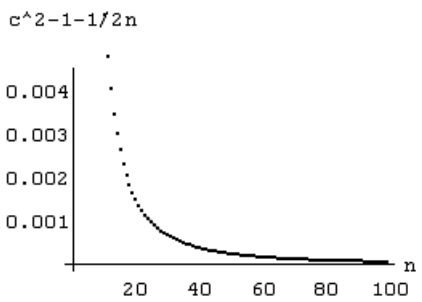
입니다. 앞서 $\tau(\theta_1,\theta_2)$에 대한 MVUE가 $U=\overline{X}+z_\alpha c_n S$인 것을 보였는바

$$Var(U;\theta_1,\theta_2) = \frac{\theta_2}{n} + z_\alpha^2(c_n^2-1)\theta_2$$

이므로 (여기서 $c_n=\sqrt{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)/\Gamma\left(\frac{n}{2}\right)$), 크래머-라오 부등식이 옳다면

$$c_n^2 - 1 - \frac{1}{2n} \geq 0, \quad \text{for } n \geq 2$$

이여야 합니다. 정말 그럴까요? 이 함수를 플롯한 <그림 6>을 보십시오.



<그림 6> $\left\{c_n^2-1-\frac{1}{2n};\ n=2,\cdots,100\right\}$의 플롯

5.5 MLE(최대가능도추정) 이론

앞 절에서, 포아송분포 Poisson(θ)로부터의 임의표본 $X_1, \cdots, X_n$에서 θ에 대한 MVUE는 $\overline{X}$이지만 $e^{-\theta}$에 대한 MVUE는 $e^{-\overline{X}}$가 아닌 $(1-\frac{1}{n})^{n\overline{X}}$였습니다. 번거롭기도 하고 뭔가 자연스럽지가 않습니다. 그러면 4장에서 소개한 최대가능도추정(maximum likelihood estimation; MLE)은 이 경우 어떤 추정치를 내놓는지 보기로 합시다. 가능도와 로그 가능도가 각각

$$L(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!} \propto e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i}, \quad \theta > 0,$$

$$l(\theta) = \log_e L(\theta; x_1, \cdots, x_n) = -n\theta + \sum_{i=1}^{n} x_i \cdot \log\theta, \quad \theta > 0$$

이므로

$$\frac{\partial}{\partial\theta} l(\theta; x_1, \cdots, x_n) = -n + \frac{\sum_{i=1}^{n} x_i}{\theta} = 0$$

로부터 $\hat{\theta} = \overline{X}$입니다. 그러면 $e^{-\theta}$에 대한 최대가능도추정치는 무엇일까요?

$\tau = e^{-\theta}$로 놓으면 (즉 $\theta = -\log_e\tau$, $0 < \tau < 1$), 로그 가능도 $l(\theta)$가

$$l^*(\tau) = l(-\log_e\tau) = -n(-\log_e\tau) + \sum_{i=1}^{n} x_i \cdot \log(-\log_e\tau), \quad 0 < \tau < 1$$

로 재표현됩니다. 따라서

$$\frac{\partial}{\partial\tau} l^*(\tau) = \frac{n}{\tau} + \sum_{i=1}^{n} x_i \cdot \frac{1}{-\log_e\tau} \cdot -\frac{1}{\tau}$$

가 됩니다. 따라서

$$\frac{\partial}{\partial\tau} l^*(\tau) = 0 \Rightarrow \log_e\hat{\tau} = -\frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \hat{\tau} = e^{-\overline{x}}$$

를 얻습니다. 즉 $e^{-\theta}$에 대한 최대가능도추정치는 단순히 $e^{-\overline{X}}$입니다. 일반적으로 이와 같은 성질을 불변성(不變性, invariance property)이라고 하는데, MVUE(최소분산비편향추정)와 달리, MLE(최대가능도추정)는 일반적으로 이런 성질을 갖습니다.

최대가능도추정치의 불변성(invariance property of MLE)

$\hat{\theta}$을 θ에 대한 최대가능도추정치(maximum likelihood estimate; mle)라고 하면 $\tau(\theta)$에 대한 mle는 $\tau(\hat{\theta})$입니다 (여기서 $\tau(\theta)$는 θ의 함수).

증명은 약간 까다롭습니다. $\tau = \tau(\theta)$로 파라미터를 θ에서 τ로 바꿈에

따라 가능도 함수 $L(\theta)$로부터 새로운 가능도 함수

$$L^*(\tau) = \max_{\theta:\tau(\theta)=\tau} L(\theta)$$

가 생성됩니다. 따라서

$$L^*(\hat{\tau}) = \max_{\tau} \max_{\theta:\tau(\theta)=\tau} L(\theta) = \max_{\theta} L(\theta) = L(\hat{\theta})$$

인데,

$$L(\hat{\theta}) = \max_{\theta:\tau(\theta)=\tau(\hat{\theta})} L(\theta) = L^*(\tau(\hat{\theta}))$$

가 됩니다. 따라서 $\tau(\hat{\theta})$가 τ에 대한 mle입니다. ■

재표현된 파라미터를 추정할 때 추정치의 재표현 값을 그대로 써도 된다는 것은 분석자의 마음을 편하게 해줍니다. MLE는 이런 점에서 좋습니다. MLE의 좋은 점 또 하나는 최대가능도추정량(maximum likelihood estimator; mle)이 충분통계량으로 표현된다는 점입니다. (앞서 보았듯이 충분통계량으로 표현되지 않는 추정량은 단순히 충분통계량에 의한 조건부 기대값에 의하여 개량될 수 있습니다. 5.4절의 라오-블랙웰 정리 참조).

최대가능도추정량과 충분통계량

$X_1, \cdots, X_n$을 확률(밀도)함수 $f(x;\theta)$로부터의 임의표본이라고 하고 S가 θ를 위한 충분통계량이라고 합시다. 그러면 θ에 대한 mle $\hat{\theta}$은 충분통계값 s의 함수로 표현됩니다.

왜냐하면 θ에 대한 가능도 함수가

$$L(\theta; x_1, \cdots, x_n) = g(S(x_1, \cdots, x_n); \theta)\, h(x_1, \cdots, x_n)$$

의 형태를 취하므로 $L(\theta; x_1, \cdots, x_n)$을 θ에 대하여 최대화하는 것과

$$g(s;\theta), \quad \text{여기서 } s = S(x_1, \cdots, x_n)$$

을 θ에 대하여 최대화하는 것은 마찬가지입니다. 따라서 $\hat{\theta}$은 s로 표현됩니다. ■

예를 들어, 포아송분포 Poisson(θ)로부터의 임의표본 $X_1, \cdots, X_n$에서 θ에 대한 충분통계량은 $\sum_{i=1}^{n} X_i$ 또는 $\overline{X}$입니다. 그런데 $e^{-\theta}$에 대한 mle는 $e^{-\overline{X}}$입니다. 따라서 최대가능도추정량은 '라오-블랙웰'과 같은 도구로는 쉽게 개량되지 않습니다. 그 만큼 이미 상당 수준에 와 있다는 것이지요.

이런 것 이외에, 최대가능도추정(MLE)의 더 큰 자랑은 점근적 성질이 알려져 있

고 그 성질들이 썩 좋다는 데 있습니다. 즉 표본크기 n이 '상당히' 큰 경우 mle $\hat{\theta}$의 확률적 행태(probabilistic behavior)가 잘 알려져 있다는 것입니다. 구체적으로

- mle $\hat{\theta}$의 일치성(consistency),
- mle $\hat{\theta}$의 점근적 정규성(asymptotic normality)
- mle $\hat{\theta}$의 점근적 비편향성(asymptotic unbiasedness)
- mle $\hat{\theta}$의 점근적 효율성(asymptotic efficiency)

등 입니다. 이제부터 $X_1, \cdots, X_n$을 확률(밀도)함수 $f(x;\theta)$, $\theta \in \Theta$로부터의 임의표본이라고 하겠습니다. 그러면

$$l(\theta; X_1, \cdots, X_n) = \sum_{i=1}^{n} \log_e f(X_i;\theta)$$

가 로그 가능도입니다. 여기서 θ에 대한 mle $\hat{\theta}$이

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n} \log_e f(X_i;\theta) = 0$$

의 근(根)이라고 가정하겠습니다.

mle $\hat{\theta}$의 일치성(consistency) :

몇 가지 적절한 정칙조건 하에서, 표본크기 n이 무한히 커짐에 따라 mle $\hat{\theta}$은 θ에 확률적으로 수렴합니다 : Under appropriate regularity conditions, as n goes to infinity, mle $\hat{\theta}$ converges to θ in probability.

정식 증명은 꽤 까다롭기 때문에 여기서는 기본 아이디어를 풀어쓰기로 합니다. 대수의 법칙에 따라

$$\frac{1}{n}\sum_{i=1}^{n} \log_e f(X_i;\theta_0) \;\rightarrow\; E\{\log_e f(X_1;\theta_0);\theta\}, \text{ in probability}$$

입니다 (여기서 θ_0은 임의의 파라미터이고 θ이 임의표본을 생성시킨 참 파라미터를 표기함). 그런데

$$\hat{\theta} = \text{argmax}_{\theta_0} \frac{1}{n}\sum_{i=1}^{n} \log_e f(X_i;\theta_0), \quad (\because\ \hat{\theta}\text{의 정의에 따라})$$

$$\theta = \text{argmax}_{\theta_0}\, E\{\log_e f(X_1;\theta_0);\theta\} \quad (\because\ \text{연습문제 5.14})$$

입니다 (여기서 'arg'는 argument를 뜻함). 따라서 $\hat{\theta}$은 θ에 확률적으로 수렴하게 됩니다 :

$$\hat{\theta} \;\rightarrow\; \theta, \qquad \text{in probability.} \qquad \blacksquare$$

mle $\hat{\theta}$의 점근적 정규성(asymptotic normality) :

몇 가지 적절한 정칙조건 하에서, 큰 표본크기 n의 임의표본으로부터의 mle $\hat{\theta}$은 근사적으로 평균 θ, 분산 $\{n I_1(\theta)\}^{-1}$인 정규분포를 따릅니다 : Under appropriate regularity conditions, mle $\hat{\theta}$ is distributed approximately to the normal distribution $\mathrm{N}(\theta, \{n I_1(\theta)\}^{-1})$ for large n.

그 논리는 대충 다음과 같습니다. 테일러 1차 근사에 의하여

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log_e f(X_i;\hat{\theta}) - \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log_e f(X_i;\theta)$$

$$= \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log_e f(X_i;\theta_1)(\hat{\theta}-\theta)$$

인데 (여기서 θ_1은 θ와 $\hat{\theta}$ 사이의 적절한 실수임), 좌변의 첫 항은 0이므로 ($\because$ mle $\hat{\theta}$에 대한 가정에 의하여),

$$\sqrt{n}(\hat{\theta}-\theta) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log_e f(X_i;\theta)}{-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log_e f(X_i;\theta_1)}$$

이 됩니다. 그런데 $n \to \infty$에 따라

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log_e f(X_i;\theta) \to \mathrm{N}(0, I_1(\theta)) \quad \text{in distribution by CLT}$$

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log_e f(X_i;\theta_1) \to I_1(\theta) \quad \text{in probability by WLLN and the consistency of mle } \hat{\theta}$$

이므로

$$\sqrt{n}(\hat{\theta}-\theta) \to \mathrm{N}\left(0, \frac{1}{I_1(\theta)}\right), \text{ in distribution}$$

임을 알 수 있습니다.

또한, 이를 이용하여

$$\tau(\hat{\theta}) - \tau(\theta) \simeq \tau'(\theta)(\hat{\theta}-\theta)$$

이므로 $\tau(\theta)$에 대한 mle $\tau(\hat{\theta})$은 근사적으로 $\mathrm{N}\left(\tau(\theta), \frac{\{\tau'(\theta)\}^2}{n I_1(\theta)}\right)$를 따름을 알 수 있습니다. θ가 벡터인 경우로 확장하면 다음과 같습니다.

$$\tau(\hat{\boldsymbol{\theta}}) \sim \mathrm{N}\left(\tau(\boldsymbol{\theta}), n^{-1} \nabla\tau(\boldsymbol{\theta})^t \{I_1(\boldsymbol{\theta})\}^{-1} \nabla\tau(\boldsymbol{\theta})\right), \quad \text{근사적으로.} \quad \blacksquare$$

앞의 결과에서 $\hat{\theta}$의 평균이 근사적으로 θ이라는 것은 mle의 점근적 비편향성을, $\hat{\theta}$의 분산이 근사적으로 $\{ n I_1(\theta) \}^{-1}$, 즉 크래머-라오 하한(下限, lower bound)이 된다는 것은 mle의 점근적 효율성을 의미하는 것입니다. 따라서 큰 표본에서는 θ에 대한 추정치로서 mle $\hat{\theta}$을 좋아하지 않을 이유가 없습니다.

한 예로서, 포아송분포 Poisson(θ)로부터의 임의표본 $X_1, \cdots, X_n$에서 θ에 대한 mle는 $\overline{X}$이고 이것의 표집분포는 근사적으로 $\mathrm{N}\left(\theta, \frac{\theta}{n}\right)$입니다. (참고: $\overline{X}$의 정확한 표집분포는 n^{-1} Poisson($n\theta$)입니다). 수치적으로 $\theta = 1$이고 $n = 16$인 경우

$$\text{근사분포 } \overline{X} \sim \mathrm{N}(1, (0.25)^2) : P\{\overline{\mathrm{X}} \leq 1.4\} = 0.9452,$$

$$\text{정확분포 } \sum X_i \sim \mathrm{Poisson}(16) : P\{\overline{\mathrm{X}} \leq 1.4\} = 0.9418$$

로서 근사값이 꽤 그럴듯하게 맞습니다.

다음은 좀 더 본격적인 사례입니다. $X_1, \cdots, X_n$을 확률밀도함수

$$f(x;\theta) = \frac{1}{2}(1 + \theta x), \quad -1 \leq x \leq 1, \quad -1 \leq \theta \leq 1 \tag{5}$$

로부터의 임의표본이라고 가정하겠습니다. 여기서 θ에 대한 mle $\hat{\theta}$을 구해보기로 합시다. 로그 가능도가

$$l(\theta) = \sum_{i=1}^{n} \log_e (1 + \theta x_i) - n \log_e 2$$

이므로

$$\frac{\partial}{\partial \theta} l(\theta) = \sum_{i=1}^{n} \frac{x_i}{1 + \theta x_i}, \quad \frac{\partial^2}{\partial \theta^2} l(\theta) = -\sum_{i=1}^{n} \left\{ \frac{x_i}{1 + \theta x_i} \right\}^2$$

입니다. 따라서 뉴튼-라프슨 방법 등으로 mle $\hat{\theta}$을 구할 수 있습니다 (4.2절 참조). 그런데 한 가지 조심할 것은 파라미터 θ가 반드시 -1과 1 사이여야 한다는 것입니다. 따라서, 간혹, 로그 가능도의 최대점이 $\tilde{\theta} \geq 1$에서 나오면 mle $\hat{\theta}$는 1이고 $\tilde{\theta} \leq -1$에서 나오면 $\hat{\theta}$는 -1인 것입니다. 그러므로 이런 경우엔 mle $\hat{\theta}$이

$$\frac{\partial}{\partial \theta} l(\theta) = \sum_{i=1}^{n} \frac{x_i}{1 + \theta x_i} = 0$$

의 근이지 않습니다. 그러나 n이 커질수록 이런 예외적 상황은 거의 발생하지 않을 것으로 예상할 수 있겠습니다.

mle $\hat{\theta}$의 분포를 수치적으로 알아보기 위하여 몬테칼로 시행을 1,000번 해보기로 합시다. 그러면 $\hat{\theta}$의 실현값을 1,000개 얻을 수 있을 테니까 경험적 분포(즉, 히스토

<표 3> mle $\hat{\theta}$의 모의분포를 만들기 위한 SAS 프로그램 : $\theta = 0.5, n = 40$

```
/* Simulation of Angular Observations and MLE's */
/* angular.iml                                  */

proc iml;
   Nrepeat = 1000;   ttheta = 0.5;    n = 20;       /*ttheta = true theta value */
   MLE = j(Nrepeat, 1, 0);

   do repeat = 1 to Nrepeat;
      ncount = 1;  x = j(n,1,0);
      do while (ncount <= n);
         xtemp = 2*uniform(0)-1;   y = uniform(0);   ytemp = (1+ttheta*xtemp)/2;
         if y <= ytemp then do;  x[ncount] = xtemp;  ncount = ncount + 1;  end;
      end;

     theta = 0;                                    /* user-supplied initial value     */
     maxtol = 0.0001;  maxiter = 10000;

     start ell;
        lik = 0;  lik1 = 0;  lik2 = 0;
        do i=1 to n;
           lik1 = lik1 + x[i]/(1+theta*x[i]);      /* first derivative  of log-lik */
           lik2 = lik2 - (x[i]/(1+theta*x[i]))**2; /* second derivative of log-lik */
        end;
     finish;

     iter = 0;
     tol  = 1;

     do while (iter <= maxiter & tol > maxtol);
        run ell;
        theta1 = theta - lik1/lik2;
        tol = abs(theta1 - theta);
        theta = theta1;
        m = 3*sum(x)/n;
        iter = iter + 1;
     end;

     if theta <= -1 then theta = -1;
        if theta >=  1 then theta = 1;
     MLE[repeat,1] = theta;
     end;

   mean = sum(MLE) / Nrepeat;
   variance = (ssq(MLE) - Nrepeat*mean*mean)/(Nrepeat-1);

   print MLE;
   print Nrepeat n ttheta mean variance;
quit;
```

그램)를 그려볼 수 있을 것입니다. 그것을 MLE의 점근적 이론에 따른 정규분포와 비교해 보기로 합시다. <표 3>은 $\theta = 0.5$로 고정시켜놓고 $n = 40$으로 하여 $\hat{\theta}$를 구하는 작업을 독립적으로 1,000번 반복하는 SAS 프로그램입니다.

이 임의시행 연구의 단계는 다음과 같습니다.

1) 모분포 $f(x;\theta)$로부터 n개의 iid 표본값을 만들어냅니다.
2) 그 표본으로부터 mle $\hat{\theta}$을 구합니다 (뉴튼-라프슨 방법).
3) 단계 1과 2를 N번 반복하여 $\hat{\theta}$의 실현값 $\hat{\theta}_1, \cdots, \hat{\theta}_N$을 확보합니다.
4) 그 값들의 분포를 그립니다.

여기서 단계 1만 구체적으로 설명하기로 하겠습니다. (5)의 분포는 이제까지 다루었던 분포가 아니니까요. 확률밀도함수

$$f(x;\theta) = \frac{1}{2}(1+\theta x), \quad -1 \le x \le 1, \quad -1 \le \theta \le 1$$

는 다행스럽게 유한구간에서 정의되어 있고 최대밀도가 유한합니다. 이런 경우엔 소위 채택-기각 방법으로 확률변수를 쉽게 실현시킬 수 있습니다.

채택-기각 방법(acceptance-rejection method) :

확률밀도함수 $f(x;\theta)$가 유한구간 $a \le x \le b$에서만 양의 값을 취하고 $\max_x f(x;\theta) \le c$라고 합시다 $(0 < c < \infty)$. U_1, U_2를 독립적으로 균일분포 Uniform(0,1)에서 발생시켜

$$X := a + (b-a)\,U_1, \quad Y := c\,U_2$$

로 놓되

$$Y \le f(X;\theta)$$

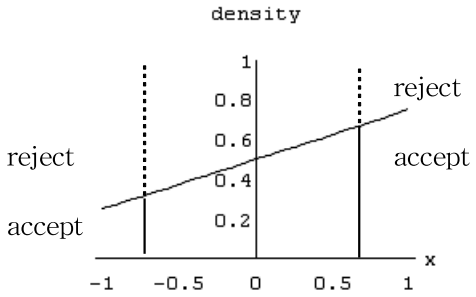
이면 X를 살리고(채택하고) 그렇지 않으면 X를 죽이면(기각하면), X는 확률밀도 $f(x;\theta)$, $a \le x \le b$를 따르게 됩니다. <그림 7>을 보십시오. ■

모의시행으로 구한 1,000개의 mle $\hat{\theta}$들의 평균과 분산은 다음과 같습니다:

평균 0.4988, 분산 0.0625.

그리고 <그림 8>은 이들 값을 히스토그램으로 만든 결과입니다. MLE에 관한 점근적 이론에 따라 mle $\hat{\theta}$은 평균과 분산이 각각

$$\text{평균 } \theta \ (= 0.5), \quad \text{분산 } \frac{1}{n\,I_1(\theta)} \ (= 0.0639)$$

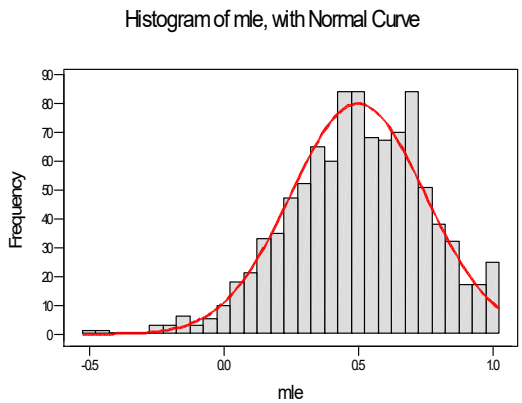


<그림 7> 확률밀도 (5)와 채택-기각 방법의 원리

인 정규분포를 따르게 되어 있습니다. 왜냐하면

$$I_1(\theta) = \int_{-1}^{1} \frac{x_1^2}{(1+\theta x_1)^2} \frac{1+\theta x_1}{2}\, dx_1 = \begin{cases} \frac{1}{3}, & \text{if } \theta = 0, \\ \frac{1}{2\theta^3}\left(\log_e \frac{1+\theta}{1-\theta} - 2\theta\right), & \text{if } \theta \neq 0 \end{cases}$$

이기 때문입니다. 그러므로 모의시행 결과와 MLE의 점근적 이론의 적용결과는, mle $\hat{\theta}$은 평균과 분산의 측면에서, 거의 같음을 볼 수 있습니다. 그러나 속 내용에 있어서는 약간의 차이가 있다고 하겠는데 특히 $\hat{\theta} = 1$에 어느 정도의 밀도가 뭉쳐있음을 볼 수 있습니다.



<그림 8> mle $\hat{\theta}$의 모의분포 히스토그램 (정규곡선 추가)

5.A 연습문제

5.1 $X_1, \cdots, X_n$을 Uniform$(\theta-\alpha, \theta+\alpha)$로부터의 iid 확률변수라고 합시다. θ를 위한 추정량으로

$$T_1 = \overline{X} \quad \text{또는} \quad T_2 = \frac{1}{2}\{\min(X_1, \cdots, X_n) + \max(X_1, \cdots, X_n)\}$$

을 고려하기로 합시다. 비편향성과 변동성의 관점에서 T_1과 T_2를 비교해보세요.

5.2 $X_1, \cdots, X_n$을 Exponential$(\theta, 1)$로부터의 iid 확률변수라고 합시다. 즉,

$$f_X(x;\theta, 1) = \exp\{-(x-\theta)\}, \quad x \geq \theta$$

입니다 (4.5절 참조). 이 때, θ를 위한 추정량으로

$$T_1 = \overline{X} - k_1 \quad \text{또는} \quad T_2 = \min(X_1, \cdots, X_n) - k_2$$

를 생각하기로 합시다. T_1과 T_2가 θ에 대한 비편향추정량이 되도록 상수 k_1과 k_2를 정하세요. 그리고 변동성의 관점에서 T_1과 T_2를 비교해보세요.

5.3 $X_1, \cdots, X_9$을 Exponential(1)로부터의 9개 iid 확률변수라고 합시다. 중위수 T의 기대값을 몬테칼로 방법으로 계산해보세요. 답은 5.1절 안에 있습니다.

5.4 $X_1, \cdots, X_n$이 정규분포 N(θ_1, θ_2)로부터의 iid 확률변수라고 합시다. $\theta_2 (= \sigma^2)$에 대한 추정량 $T_1 = \sum_{i=1}^{n} (X_i - \overline{X})^2 / (n-1)$과 $T_2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 / n$의 MSE(평균제곱오차)가

$$\text{MSE}(T_1, \theta_1, \theta_2) = \theta_2^2 \cdot \frac{2}{n-1}, \qquad \text{MSE}(T_2, \theta_1, \theta_2) = \theta_2^2 \cdot \frac{2n-1}{n^2}$$

임을 보이세요 (5.3절 참조).

5.5 (계속) T_1과 T_2의 MSLE 값을 수치적분을 사용하여 구해보세요 (5.3절 참조).

5.6 (계속) $\theta_2 (= \sigma^2)$에 대한 추정량으로 $T = \sum_{i=1}^{n} (X_i - \overline{X})^2 / k_n$을 고려하기로 합시다 (여기서 k_n은 상수). 그리고 MSE에 대한 또 다른 대안으로

$$\text{MSGE}(T, \theta_1, \theta_2) = E\left\{\frac{T}{\theta_2} - 1 - \log_e \frac{T}{\theta_2}\ ; \theta_1, \theta_2\right\}$$

를 생각합시다. t의 함수인 $G_e(t;\theta_2) = \frac{t}{\theta_2} - 1 - \log_e \frac{t}{\theta_2}$를 플롯해보세요. MSGE 기준을 최소로 하는 상수 k_n을 구하시오.

답. $k_n = n-1$.

5.7 다음을 증명하세요.

$$Var(T) = E\{Var(T \mid S)\} + Var\{E(T \mid S)\}.$$

5.8 (X, Y)가 이변량정규분포 $\mathrm{BN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$를 따르는 경우, $Y = y$가 주어졌을 때 X의 조건부 분포가

$$\mathrm{N}\left(\mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2), (1 - \rho^2)\sigma_1^2\right)$$

임을 보이세요.

5.9 $X_1, \cdots, X_n$이 정규분포 $\mathrm{N}(\theta_1, \theta_2)$로부터의 임의표본인 경우, $c_n S$의 기대값이 모표준편차 $\sqrt{\theta_2}$가 되도록 상수 c_n을 정하세요.

$$\text{답 : } c_n = \sqrt{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)/\Gamma\left(\frac{n}{2}\right).$$

5.10 크기 n의 iid 임의표본에서, 피셔 정보량 $I_n(\theta)$은 단위 정보량 $I_1(\theta)$의 n 배임을 증명하세요.

5.11 Poisson(θ) 임의표본 $X_1, \cdots, X_n$으로 $\tau(\theta) = e^{-\theta}$에 대한 비편향추정량을 생각하기로 합시다 (5.4절 참조). 이 때, $e^{-\theta}$에 대한 MVUE에 의하여도 크래머-라오 부등식의 하한이 달성될 수 없음을 보이세요.

5.12 포아송 분포, 베르누이 분포, 지수분포, 감마분포, 정규분포에 대한 단위 피셔 정보량을 구하세요.

5.13 $X_1, \cdots, X_n$을 지수분포 Exponential(θ)로부터의 임의표본이라고 할 때, $\tau(\theta) = e^{-c/\theta}$ $(= P\{\mathrm{X}_1 \geq \mathrm{c}; \theta\})$, (c는 상수)에 대한 MVUE와 MLE를 구해보세요.

5.14 적절한 정칙조건하에서 $E\{\log_e f(X_1; \theta_0); \theta\}$가 θ_0의 함수로서 $\theta_0 = \theta$일 때 최대가 됨을 증명하세요.

5.15 $X_1, \cdots, X_n$을 확률밀도함수

$$f(x; \theta) = \frac{1}{2}(1 + \theta x), \quad -1 \leq x \leq 1, \quad -1 \leq \theta \leq 1$$

로부터의 임의표본이라고 가정하고 (5.5절의 계속) $\theta = 0.0$과 0.5, $n = 40$과 100일 때, mle $\hat{\theta}$을 1,000개 확보하여 평균과 분산 및 히스토그램을 구하고 점근적 이론의 결과와 비교해보세요.

탐구문제 : $X_1, \cdots, X_n$이 정규분포 $N(\theta, \sigma^2)$로부터의 임의표본인 경우(σ^2는 旣知)에서 $\tau(\theta) = \Phi\left(\frac{c-\theta}{\sigma}\right)$에 대한 MVUE

$$\phi_T(\overline{X}) = \Phi\left(\sqrt{\frac{n}{n-1}} \cdot \frac{c-\overline{X}}{\sigma}\right)$$

의 분산을 구해보세요. [힌트: 정확한 분산은 구하기 어렵습니다. 델타방법을 사용하여 근사적 계산을 하든가 몬테칼로 방법을 사용하여 분산을 통계적으로 추정해보십시오).

5.B 읽을만한 책

점 추정론에 대하여는 일반적인 수리통계학 책들이 잘 다루고 있습니다. 다음 중에서 아무 책이나 보기 바랍니다.

- Hogg, R.V. and Craig, A.T. (1995) *Introduction to Mathematical Statistics*, 5th Edition. Prentice Hall. (Chapters 6 and 8)
- Bickel, P.J. and Duksum, K.A. (1977) *Mathematical Statistics*.. Holden-Day. (Chapters 3 and 4)
- Rice, J.A. (1995) *Mathematical Statistics and Data Analysis*, 2nd Edition. Duxbury Press. (Chapter 8)
- Casella, G. and Berger, R.L. (1990) *Statistical Inference*. Duxbury. (Chapter 7)
- Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. Chapman and Hall. (Chapter 8 and Section 9.2)

수치적분과 채택-기각 방법에 더 관심이 있다면 다음 문헌을 보기 바랍니다.

- 최영훈・이승천 (1995)「C에 의한 전산통계」 자유 아카데미. (4장, 5장)
- Burden, R.L. and Faires, J.D. (1997) *Numerical Analysis*, 6th Edition. Brooks and Cole. (Chapter 4)