7장. 구간 추정론

5장에서 다루었던 점 추정이 파라미터 θ를 점 찍듯이 맞추는 것이라면 이 장에서는 구간으로 파라미터 θ를 담아내는 방법을 다루게 됩니다. 비유적으로 이야기하자면, 물고기를 잡는데 창으로 찍는 것이 점 추정이고 그물에 걸리게 하는 것은 구간추정입니다.

통계적 구간을 얻기 위해서 추축량(pivotal quantity)을 활용하는 방법과 양측 가설검증을 활용하는 방법이 있습니다. 이 두 방법은 소표본에서 각 경우마다 개별적으로 정교하게 만들어집니다.

그러나, 대표본에서는 구간추정을 하는 일반적인 방법이 있습니다. 최대가능도 추정 이론 또는 일반화 가능도 비에 관한 이론에 근거하여 근사적 신뢰구간을 만들어 낼 수 있답니다. 이 때, 그 근사적 구간이 실제 어느 정도의 상대적 빈도로 미지의 파라미터인 θ를 담아내는지에 대하여는 확인이 필요합니다.

가설검증에서와 마찬가지로 구간추정에서도 비편향성의 개념이 있고 비편향적 구간들 중에서 가장 좋은 구간이 어떤 구간이냐를 따지는 최적성 이론이 있습니다. 그리고 마지막으로 신뢰구간과 유사한 예측구간에 대하여 간단한 설명이 있겠습니다.

이론적 측면에서는 가설검증과 신뢰구간은 동전의 양면이라고 할 수도 있습니다. 그러나 통계적 역할 면에서는 확연히 다른데 '가설검증'이 제시된 가설의 확증을 목표로 한다면 '구간추정'은 자료 요약적 성격이 강합니다. 물론, 구간추정은 단순한 기술적 요약과는 다릅니다. 구간추정은 표본의 불확정성을 염두에 넣은 추론적 요약이라고 할까요.

차례 :

7.1 추축량(Pivotal Quantity)을 활용하는 방법

$X_1, \cdots, X_n$을 정규분포 $\mathrm{N}(\theta, \sigma_0^2)$으로부터의 임의표본이라고 합시다 ($\sigma_0^2$은 미리 알려짐). 그러면 잘 알려진 대로

$$\frac{\overline{X} - \theta}{\sigma_0 / \sqrt{n}} \sim \mathrm{N}(0,1) \tag{1}$$

을 따릅니다. 따라서

$$P\left\{ -z_{\alpha/2} \leqq \frac{\overline{X} - \theta}{\sigma_0 / \sqrt{n}} \leqq z_{\alpha/2} \mid \theta \right\} = 1 - \alpha$$

가 됩니다. 이것을 재표현하면

$$P\left\{ \overline{X} - z_{\alpha/2}\sigma_0 / \sqrt{n} \leqq \theta \leqq \overline{X} + z_{\alpha/2}\sigma_0 / \sqrt{n} \mid \theta \right\} = 1 - \alpha$$

입니다. 즉

$$\left(\overline{X} - z_{\alpha/2}\sigma_0 / \sqrt{n} \, , \, \overline{X} + z_{\alpha/2}\sigma_0 / \sqrt{n} \right) \tag{2}$$

이라는 임의구간을 사용함으로써 θ가 걸려들게 할 수 있는 것이지요. 그러나 낚시가 그렇듯이 항상 성공할 수 있는 것은 아닙니다. 확률 $1-\alpha$로만 성공할 수 있는데, 통상 $1-\alpha$를 95%나 90%로 놓습니다. 우리는 이 확률을 **신뢰수준**(confidence level)이라고 하고 파라미터 θ를 낚는 데 사용한 임의구간을 **신뢰구간**(confidence interval)이라고 합니다.

앞의 예에서 실제 관측 표본값을 $x_1, \cdots, x_n$이라고 하면 θ에 대한 수준 $1-\alpha$의 신뢰구간은

$$\left(\overline{x} - z_{\alpha/2}\sigma_0 / \sqrt{n} \, , \, \overline{x} + z_{\alpha/2}\sigma_0 / \sqrt{n} \right) \tag{3}$$

으로 산출됩니다. 그러나, 이 구간은 임의구간(random interval)이 아니기 때문에

$$P\left\{ \theta \in \left(\overline{x} - z_{\alpha/2}\sigma_0 / \sqrt{n} \, , \, \overline{x} + z_{\alpha/2}\sigma_0 / \sqrt{n} \right) \right\} = 1 - \alpha$$

라고 할 수 없습니다. 그러므로 신뢰수준 $1-\alpha$는 (3)의 구간에 붙는 확률이 아니라 (2)와 같은 구간들에 부여되는 확률을 의미합니다. 이 정도는 통계학 입문 과목에서 깨치고 왔겠지요? [그러나, 이것은 통계학계에서 주류(主流)인 빈도학파(frequentist school)의 해석일 뿐입니다. 다른 해석도 가능합니다. 허명회 (1997) 참조.]

앞의 예에서 θ에 관한 신뢰구간을 구하는 데 무엇이 결정적인 역할을 하였나 다시 살펴봅시다. 그것은 (1)의 임의량(random quantity)이 θ와 관계없는 확률분포를 따른다는 사실입니다.

추축량(樞軸量, pivotal quantity) : 정의

$X_1, \cdots, X_n$을 확률(밀도)함수 $f(x;\theta)$로부터의 iid 확률변수라고 합시다. 만약 $Q(X_1, \cdots, X_n;\theta)$이 θ와 관계없는 분포를 따른다면, 임의량 Q를 추축량이라고 합니다.

앞의 예에서는 (1)이 추축량입니다. 다른 예를 들어볼까요? $X_1, \cdots, X_n$을 균일분포 Uniform$(0,\theta)$로부터의 임의표본이라고 합시다. 그러면

$$Y = \max(X_1, \cdots, X_n)$$

은 확률밀도함수

$$g(y;\theta) = n\left(\frac{y}{\theta}\right)^{n-1}\frac{1}{\theta}, \quad 0 \le y \le \theta$$

를 따릅니다. 그러므로

$$Q = \frac{Y}{\theta}$$

의 확률밀도함수는

$$h(q) = n q^{n-1}, \ 0 \le q \le 1$$

로 θ와 전혀 관계없습니다. 따라서 Q는 추축량입니다 (Q는 통계량이 아니므로, '추축통계량'이라고 하면 안 됩니다).

이제 θ에 관한 $1-\alpha$ 수준의 신뢰구간을 구해보도록 합시다.

$$\int_a^b h(q)\,dq \ (= b^n - a^n) = 1-\alpha, \quad 0 \le a \le b \le 1$$

이 되게 하기 위한 a와 b의 0과 1 사이 실수 쌍은 무수히 많겠으나 Q의 밀도함수 $h(q)$가 증가함수이므로 $b=1$로 고정시키는 것이 합당할 것입니다. 그러면

$$a = \alpha^{1/n}$$

이 됩니다. 즉,

$$1-\alpha = P\left\{\alpha^{1/n} \le \frac{Y}{\theta} \le 1\right\} = P\left\{Y \le \theta \le \frac{Y}{\alpha^{1/n}}\right\}$$

입니다. 따라서 θ에 관한 수준 $1-\alpha$ 수준의 신뢰구간은

$$\left(Y, \left(\frac{1}{\alpha}\right)^{1/n} Y\right)$$

로 주어집니다. 예컨대 $n=10$일 때, 95% 수준의 신뢰구간은 $(Y, 1.349\,Y)$입니다. $n=25$일 때 신뢰구간은 $(Y, 1.127\,Y)$입니다. 여기서 $Y = \max(X_1, \cdots, X_n)$.

추축량을 이용하여 신뢰구간을 구하는 예를 하나 더 들어보겠습니다. 임의표본 $X_1, \cdots, X_n$이 지수분포 Exponential(θ), 즉 밀도함수

$$f_X(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, \quad x \geq 0$$

으로부터 생성되었다고 합시다. 이에 따라 $\sum_{i=1}^{n} X_i$가 감마분포 Gamma(n, θ)를 따릅니다. 때문에

$$\sum_{i=1}^{n} X_i / \theta \sim \text{Gamma}(n,1), \quad \text{즉} \quad Q \equiv \frac{2}{\theta}\sum_{i=1}^{n} X_i \sim \chi^2(2n)$$

이 됩니다. 그러므로

$$P\left\{ \chi^2_{2n,\alpha/2} \leq \frac{2}{\theta}\sum_{i=1}^{n} X_i \leq \chi^2_{2n,1-\alpha/2} \right\} = 1-\alpha$$

가 됩니다. 여기서 $\chi^2_{2n,\alpha/2}$와 $\chi^2_{2n,1-\alpha/2}$는 각각 자유도 $2n$의 카이제곱분포의 $\alpha/2$분위수와 $1-\alpha/2$ 분위수입니다. 따라서 θ에 관한 수준 $1-\alpha$의 신뢰구간으로

$$\left(\frac{2n}{\chi^2_{2n,1-\alpha/2}} \cdot \overline{X}, \frac{2n}{\chi^2_{2n,\alpha/2}} \cdot \overline{X} \right)$$

를 얻습니다. 예로서 $1-\alpha = 0.95$이고 $n = 10$인 경우, $\chi^2_{2n,\alpha/2} = 9.59$, $\chi^2_{2n,1-\alpha/2} = 34.17$이기 때문에 θ에 관한 95% 수준의 신뢰구간은 $(0.585\overline{X}, 2.09\overline{X})$입니다.

문제는 "어떻게 추축량을 찾아낼 수 있는가?" 하는 것입니다. 이에 관하여 명쾌한 답은 아직 없습니다. 단, 연속형 분포의 경우 다음 결과가 여러 경우에 통하는 것으로 알려져 있습니다.

연속분포에서의 추축량 : 보조정리

$X_1, \cdots, X_n$이 밀도함수가 $f_X(x;\theta)$이고 분포함수가 $F_X(x;\theta)$인 연속형 분포로부터의 임의표본이라고 합시다. 그러면

$$Q \equiv -2\sum_{i=1}^{n} \log_e F_X(X_i;\theta)$$

는 추축량으로서 자유도 $2n$의 카이제곱분포 $\chi^2(2n)$을 따릅니다.

증명은 어렵지 않습니다. $X_1, \cdots, X_n$의 변환

$$F_X(X_1;\theta), \cdots, F_X(X_n;\theta)$$

가 독립적으로 균일분포 Uniform(0,1)을 따르게 되기 때문이고 이에 따라

$$-\log_e F_X(X_1;\theta), \cdots, -\log_e F_X(X_n;\theta)$$

가 독립적으로 지수분포 Exponential(1)을 따르기 때문입니다. ■

바로 앞의 예에 이 보조정리를 적용해보십시오. 꼭 들어맞을 것입니다. 변형된 예로서 다음 문제를 생각해 봅시다. $X_1, \cdots, X_n$이 밀도함수

$$f_X(x;\theta) = \theta\, x^{\theta-1},\ 0 < x \le 1,\ \theta > 0$$

으로부터의 임의표본인 경우, θ에 관한 신뢰구간을 구하기 위해 필요한 추축량을 구해봅시다. 분포함수가

$$F_X(x;\theta) = x^{\theta},\ 0 < x \le 1$$

이므로, 앞의 보조정리에 따라 추축량

$$Q = -2\sum_{i=1}^{n} \log_e X_i^{\theta} = -2\theta \sum_{i=1}^{n} \log_e X_i \sim \chi^2(2n)$$

이 유도됩니다. 따라서

$$P\left\{ \chi^2_{2n,\alpha/2} \le -2\theta \sum_{i=1}^{n} \log_e X_i \le \chi^2_{2n,1-\alpha/2} \right\} = 1-\alpha$$

입니다. 즉, θ에 관한 수준 $1-\alpha$의 신뢰구간으로

$$\left(\frac{\chi^2_{2n,\alpha/2}}{-2\sum_{i=1}^{n} \log_e X_i},\ \frac{\chi^2_{2n,1-\alpha/2}}{-2\sum_{i=1}^{n} \log_e X_i} \right)$$

를 얻습니다.

파라미터가 2개인 경우에서는 추축량이 단 1개의 파라미터만을 포함하여야 파라미터에 관한 신뢰구간을 구하는 데 도움이 됩니다. 대표적인 예는 $X_1, \cdots, X_n$이 정규분포 $\mathrm{N}(\mu, \sigma^2)$으로부터의 임의표본인 경우입니다 ($\sigma^2$도 알려지지 않은 파라미터임). 모평균 μ만 포함하는 추축량은

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim \mathrm{t}(n-1)$$

입니다 (여기서 s^2은 표본분산임). 이에 따라 μ에 관한 수준 $1-\alpha$의 신뢰구간이 다음과 같이 구해집니다:

$$\left(\overline{X} - t_{n-1,\alpha/2}\, s/\sqrt{n}\ ,\ \overline{X} + t_{n-1,\alpha/2}\, s/\sqrt{n}\ \right).$$

앞의 예에서 모분산 σ^2만 포함하는 추축량은

$$\frac{(n-1)\,s^2}{\sigma^2} \sim \chi^2(n-1)$$

이며, 이에 따라 σ^2에 관한 수준 $1-\alpha$의 신뢰구간은 다음과 같이 얻어집니다:

$$\left(\frac{(n-1)}{\chi^2_{n-1,1-\alpha/2}} \cdot s^2,\ \frac{(n-1)}{\chi^2_{n-1,\alpha/2}} \cdot s^2 \right).$$

예를 들어 $1-\alpha = 0.95$ 이고 $n = 10$ 인 경우, $\chi^2_{n-1,\alpha/2} = 2.70$, $\chi^2_{n-1,1-\alpha/2} = 19.02$ 이므로 σ^2 에 관한 95% 수준의 신뢰구간은 $(0.473\,s^2, 3.33\,s^2)$ 이 됩니다. $n = 25$ 인 경우엔, $\chi^2_{n-1,\alpha/2} = 12.40$, $\chi^2_{n-1,1-\alpha/2} = 39.36$ 이므로 신뢰구간은 $(0.610\,s^2, 1.94\,s^2)$ 입니다.

신뢰구간을 구하기 위하여 추축량을 활용하는 방법은 매우 매력적이기는 하지만 여러 경우에 있어 적절한 추축량을 찾아내는 것이 쉽지 않습니다. 다음 절에서 다루게 될 예들이 그런 보기들입니다.

7.2 양측 가설검증을 활용하는 방법

$X_1, \cdots, X_n$ 을 베르누이 분포 Bernoulli(θ) 로부터의 임의표본이라고 합시다. 즉 $\overline{X}$ 는 표본비율 p 입니다. 이 때, 모비율 θ 에 관한 신뢰구간을 구해보도록 합시다. 이 경우엔 적절한 추축량이 없습니다 (아직, 알려지지 않았습니다). 이런 경우엔 6장에서 다룬 바 있는 유의성 검증 문제

$$\mathrm{H}_0 : \theta = \theta_0 \quad \text{대} \quad \mathrm{H} : \theta \neq \theta_0$$

를 수준 α 에서 생각해 보는 것입니다. θ 에 대한 충분통계량 $S = \sum_{i=1}^{n} X_i$ 에 의한 영가설 H_0 의 기각역은

$$S < c_1 \quad \text{또는} \quad S > c_2$$

입니다. 이 때, c_1 과 c_2 는

$$P\{ S < c_1 \text{ or } S > c_2 \mid \theta_0 \} \leqq \alpha,$$

마찬가지로

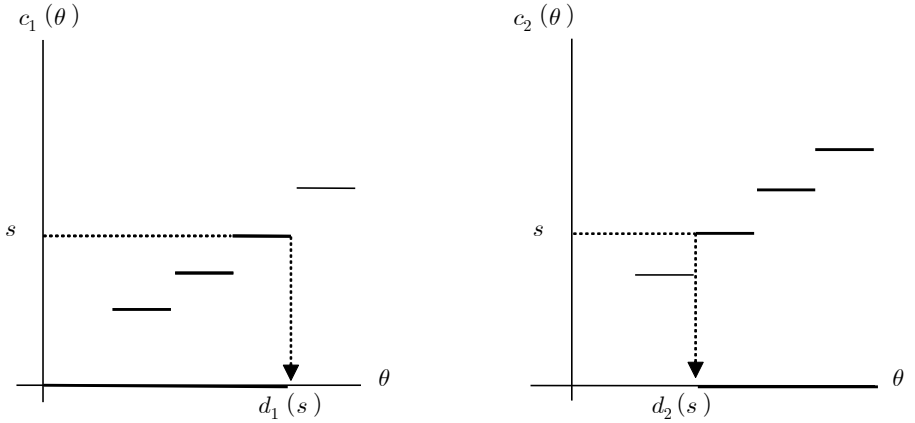
$$P\{ c_1 \leq S \leq c_2 \mid \theta_0 \} \geq 1-\alpha$$

를 만족하는 정수입니다. 예컨대 $n = 100$, $1-\alpha = 0.95$ 인 경우,

$$P\{ 31 \leq S \leq 50 \mid 0.4 \} \geq 0.95,$$
$$P\{ 40 \leq S \leq 60 \mid 0.5 \} \geq 0.95,$$
$$P\{ 50 \leq S \leq 69 \mid 0.6 \} \geq 0.95$$

입니다. 이와 같이 c_1 과 c_2 가 θ_0 에 따라 다르기 때문에 $c_1(\theta_0)$ 와 $c_2(\theta_0)$ 로 표기하는 것이 좋겠습니다. 즉,

$$P\{ c_1(\theta_0) \leq S \leq c_2(\theta_0) \mid \theta_0 \} \geq 1-\alpha. \tag{4}$$



(1) $c_1(\theta) \le s \Leftrightarrow \theta \le d_1(s)$, (2) $c_2(\theta) \ge s \Leftrightarrow \theta \ge d_2(s)$

<그림 1> 하한과 상한 함수의 뒤집기

그리고 $c_1(\theta_0)$와 $c_2(\theta_0)$가 θ_0의 증가함수가 되리라는 것은 직관적으로 쉽게 알 수 있습니다. 따라서

$$c_1(\theta_0) \le S \Leftrightarrow \theta_0 \le d_1(S), \quad S \le c_2(\theta_0) \Leftrightarrow d_2(S) \le \theta_0$$

가 되는 두 함수 $d_1(S)$와 $d_2(S)$를 생각하는 것이 가능합니다. <그림 1>을 보십시오. 예를 들어 $n = 100, 1-\alpha = 0.95$인 경우,

$$c_1(\theta_0) \le 50 \Leftrightarrow \theta_0 \le 0.6, \quad c_2(\theta_0) \ge 50 \Leftrightarrow \theta_0 \ge 0.4$$

이기 때문에 대략 $d_1(50) = 0.6$, $d_2(50) = 0.4$입니다 (더 정확한 계산은 곧 나옵니다). 따라서 (4)는

$$P\{d_2(S) \le \theta_0 \le d_1(S) \mid \theta_0\} \ge 1-\alpha,$$

로 표현가능합니다. 마찬가지로

$$P\{d_2(S) \le \theta \le d_1(S) \mid \theta\} \ge 1-\alpha$$

로 쓸 수 있게 되므로 θ에 관한 수준 $1-\alpha$의 신뢰구간으로

$$(d_2(S), d_1(S))$$

를 얻게 되는 것입니다.

<표 1> 이항표본에서 신뢰구간의 상한을 구하기 위한 SAS/IML 프로그램

```
/* Upper Confidence Limit for Binomial Case with n = 100 and s = 57 */
/* File Name: conf1.iml                                              */

proc iml;
   theta0=0;  theta1=1;  diff=1;  n = 100;  s = 57;  iter = 1;
   do while (diff > 0.0001);
      theta=(theta0+theta1)/2;
      temp1 = 1; temp2 = (1-theta)**n;
      cumprob = temp2;
      do k=1 to s;
         temp1 = temp1*(n-k+1)/k*theta/(1-theta);
         prob = temp1*temp2;
         cumprob = cumprob + prob;
      end;
      diff = abs(cumprob-0.025);
      if cumprob < 0.025 then print iter cumprob[format=8.4] theta[format=8.4];
      if cumprob > 0.025 then theta0 = theta;
         else theta1=theta;
      iter = iter + 1;
   end;
quit;
```

예를 들어 $n = 100$ 에서 $S = 57 (= s)$ 이 관측된 경우 95% 수준의 신뢰구간을 정확히 구해보기로 하겠습니다. 신뢰구간 $(d_2(s), d_1(s))$ 의 상한인 $\theta = d_1$ 에서는

$$P\{S \leq s \mid \theta\} = 0.025 \tag{5}$$

여야 할 것입니다. <그림 1>을 보십시오. 그런데

$$P\{S \leq s \mid \theta\} = \sum_{k=0}^{s} \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

이므로 일종의 누적확률 계산이 필요합니다. 이것을 수치적으로 정확히 계산하기 위해서는 각 항을 각기 산출하여 단순히 더하는 것보다는

$$\sum_{k=0}^{s} \binom{n}{k} \theta^k (1-\theta)^{n-k} = \sum_{k=0}^{s} \left\{ \prod_{j=1}^{k} \frac{n-j+1}{j} \frac{\theta}{1-\theta} \right\} (1-\theta)^n$$

을 활용하는 것이 좋습니다. 그리고 이분법(bisection method; 4.2절 참조)에 의하여 (5)를 만족하는 θ를 찾을 수 있을 것입니다. <표 1>의 SAS/IML 프로그램을 돌려 찾은 θ 값은 0.6687입니다. 즉 $d_1(57) = 0.6687$.

한편, 신뢰구간 $(d_2(s), d_1(s))$ 의 하한인 $\theta = d_2$ 에서는

$$P\{S \geq s \mid \theta\} = 0.025$$

여야 합니다. 그런데

$$P\{S \geq s \mid \theta\} = \sum_{k=s}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k}$$
$$= \sum_{n-k=0}^{n-s} \binom{n}{n-k} (1-\theta)^{n-k} \theta^{n-(n-k)}$$
$$= \sum_{k'=0}^{n-s} \binom{n}{k'} (1-\theta)^{k'} \theta^{n-k'} = P\{S \leq n-s \mid 1-\theta\}$$

이므로 <표 1>의 SAS 프로그램을 그대로 활용하여 신뢰구간의 하한을 구할 수 있습니다. 단, $s\,(=57)$ 대신에 $n-s\,(=43)$ 를 대입하여 프로그램을 수행하고 출력된 θ 값$(=0.5330)$으로부터 $1-\theta$를 산출하여야 합니다. 이렇게 하면 θ 값으로 0.4670을 얻게 됩니다. 즉 $d_2\,(57) = 0.4670$. 따라서 $n = 100$이고 $s = 57$인 경우, 성공확률 θ에 대한 수준 95%의 신뢰구간은 $(0.467, 0.669)$입니다.

이와 같이 양측 가설검증을 활용하여 신뢰구간을 구하는 예를 하나 더 해보기로 하겠습니다. $X_1, \cdots, X_n$을 포아송 분포 Poisson(θ)로부터의 임의표본이라고 합시다. θ에 관한 신뢰구간을 구해보도록 합시다.

θ를 위한 충분통계량 $S = \sum_{i=1}^{n} X_i$는 Poisson$(n\theta)$를 따릅니다. 그러므로

$$P\{S \leq c_1(\theta) - 1 \mid \theta\} < \alpha/2, \quad P\{S \geq c_2(\theta) + 1 \mid \theta\} < \alpha/2$$

가 되도록, 즉

$$P\{c_1(\theta) \leq S \leq c_2(\theta) \mid \theta\} \geq 1-\alpha$$

가 되도록 $c_1(\theta)$와 $c_2(\theta)$를 정할 필요가 있습니다. 이 함수를 뒤집어서

$$P\{d_2(S) \leq \theta \leq d_1(S) \mid \theta\} \geq 1-\alpha$$

로 표현해냅니다. 이에 따라 θ에 관한 수준 $1-\alpha$의 신뢰구간으로

$$(d_2(S),\ d_1(S))$$

를 얻습니다. 기본원리는 이항표본의 경우와 같습니다.

구체적으로 $n = 4$인 임의표본에서 S의 값으로 $22\,(=s)$를 관측하였다고 합니다. 그러면 신뢰구간 $(d_2(s),\ d_1(s))$의 상한 $\theta = d_1$과 하한 $\theta = d_2$에서 각각

$$P\{S \leq s \mid \theta\} = 0.025, \qquad P\{S \geq s \mid \theta\} = 0.025$$

이여야 합니다. 즉

$$P\{S \leq s \mid \theta\} = \sum_{k=0}^{s} e^{-n\theta} \frac{(n\theta)^k}{k!} = 0.025,$$
$$P\{S \leq s-1 \mid \theta\} = \sum_{k=0}^{s-1} e^{-n\theta} \frac{(n\theta)^k}{k!} = 0.975$$

<표 2> 포아송 표본에서 신뢰구간을 구하기 위한 SAS/IML 프로그램

```
/* Confidence Interval for Poisson Case with n = 4 and s = 22 */
/* File Name :  conf2.iml                                     */

proc iml;
   n = 4;  s = 22;

   /* for upper limit */
   ntheta0=0.01;  ntheta1=20*n;  diff=1;  iter=1;
   do while (diff > 0.0001);
      ntheta=(ntheta0+ntheta1)/2;
      temp1 = 1; temp2 = exp(-ntheta);
      cumprob = temp2;
      do k=1 to s;
         temp1 = temp1*ntheta/k;
         prob = temp1*temp2;
         cumprob = cumprob + prob;
      end;
      diff = abs(cumprob-0.025);
      theta = ntheta/n;
      if cumprob < 0.025 then
         print "upper" iter cumprob[format=8.4] theta[format=8.4];
      if cumprob > 0.025 then ntheta0 = ntheta;
         else ntheta1=ntheta;
      iter = iter + 1;
   end;

   /* for lower limit */
   ntheta0=0.01;  ntheta1=20*n;  diff=1;  iter=1;
   do while (diff > 0.0001);
      ntheta=(ntheta0+ntheta1)/2;
      temp1 = 1; temp2 = exp(-ntheta);
      cumprob = temp2;
      do k=1 to s-1;
         temp1 = temp1*ntheta/k;
         prob = temp1*temp2;
         cumprob = cumprob + prob;
      end;
      diff = abs(cumprob-0.975);
      theta = ntheta/n;
      if cumprob < 0.975 then
         print "lower" iter cumprob[format=8.4] theta[format=8.4];
      if cumprob > 0.975 then ntheta0 = ntheta;
         else ntheta1 = ntheta;
      iter = iter + 1;
   end;
quit;
```

를 만족하는 $\theta = d_1$ 과 $\theta = d_2$ 를 산출하여야 합니다. <표 2>의 SAS/IML 프로그램을 보십시오. 이렇게 얻게되는 θ의 상한 값은 8.3315, 하한 값은 3.4469 입니다. 따라서 포아송 파라미터 θ에 대한 수준 95%의 신뢰구간은 (3.45, 8.33) 입니다.

7.3 근사적 신뢰구간

앞 절에서 양측 가설 검증을 뒤집어서 θ에 관한 신뢰구간을 산출하는 방법을 설명하고, 두 사례에서 구체적인 문제풀이를 하였습니다. 그러나 그 사례들에서 표본크기 n이 큰 경우에서는 신뢰구간의 상·하한의 계산이 상당히 복잡해질 것이라는 것을 짐작할 수 있습니다. 좀 간편한 방법이 있지 않을까요?

5.5절에서 최대가능도 추정량(m.l.e.) $\hat{\theta}$의 점근적 정규성과 일치성에 대하여 이야기한 바 있습니다. 다시 기술하자면, 몇 가지 적절한 정칙조건 하에서, 큰 표본크기 n의 임의표본으로부터의 $\hat{\theta}$ 은 근사적으로 평균 θ, 분산 $\{n I_1(\theta)\}^{-1}$ 인 정규분포를 따른다는 것입니다 (점근적 정규성). 즉,

$$\sqrt{n}\,(\hat{\theta}-\theta) \quad \sim \quad \mathrm{N}\left(0, \frac{1}{I_1(\theta)}\right), \text{ 근사적으로.}$$

그리고, $\hat{\theta}$ 은 θ에 대하여 점근적으로 일치한다는 것입니다 (일치성). 때문에

$$W \;\equiv\; \sqrt{n}\, I_1(\hat{\theta})^{0.5}\,(\hat{\theta}-\theta)$$

는 점근적으로 표준정규분포 $\mathrm{N}(0,1)$을 따르게 됩니다. 그러므로 W는 근사적인 추축량입니다:

$$P\{\, -z_{\alpha/2} \;\le\; \sqrt{n}\, I_1(\hat{\theta})^{0.5}\,(\hat{\theta}-\theta) \;\le\; z_{\alpha/2} \,\} \;\simeq\; 1-\alpha\,,$$

즉

$$P\{\, \hat{\theta} - z_{\alpha/2}\, I_1(\hat{\theta})^{-0.5}/\sqrt{n} \;\le\; \theta \;\le\; \hat{\theta} + z_{\alpha/2}\, I_1(\hat{\theta})^{-0.5}/\sqrt{n} \,\} \;\simeq\; 1-\alpha\,.$$

따라서

$$(\, \hat{\theta} - z_{\alpha/2}\, I_1(\hat{\theta})^{-0.5}/\sqrt{n}\,,\, \hat{\theta} + z_{\alpha/2}\,\sqrt{n}\, I_1(\hat{\theta})^{-0.5}/\sqrt{n}\,)$$

은 θ에 관한 근사적 수준 $1-\alpha$ 의 신뢰구간입니다.

앞 절에서 들었던 두 예에 이 방법을 적용해보도록 하겠습니다. 첫 번째 예인 베르누이 표본의 경우에서 파라미터 θ에 대한 mle $\hat{\theta}$ 은 $p\,(=\overline{X})$ 이고 피셔 정보 $I_1(\theta)$ 는 $\{\theta(1-\theta)\}^{-1}$ 입니다. 따라서

$$(\, p - z_{\alpha/2}\,\sqrt{p(1-p)/n}\,,\; p + z_{\alpha/2}\,\sqrt{p(1-p)/n}\,)$$

을 θ에 관한 근사적 수준 $1-\alpha$의 신뢰구간으로 얻습니다. 수치 예로서, $n = 100$이고 $p = 0.57$인 경우 θ에 관한 근사적 95% 수준의 신뢰구간은

$$\left(0.57 - 1.96 \sqrt{0.57 \cdot 0.43/100} \, , \; 0.57 - 1.96 \sqrt{0.57 \cdot 0.43/100} \right)$$
$$= (0.473, 0.667)$$

입니다. 이것은 앞에서 얻었던 신뢰구간 $(0.467, 0.669)$와 거의 비슷합니다.

두 번째 예인 포아송 표본의 경우에서 파라미터 θ에 대한 mle $\hat{\theta}$은 $\overline{X}$이고 피셔 정보 $I_1(\theta)$는 θ^{-1}입니다. 따라서

$$\left(\overline{X} - z_{\alpha/2} \sqrt{\overline{X}/n} \, , \overline{X} + z_{\alpha/2} \sqrt{\overline{X}/n} \right)$$

이 θ에 관한 근사적 수준 $1-\alpha$의 신뢰구간입니다. 따라서 $n = 4$이고 $\overline{X} = 5.5$인 표본에서 θ에 관한 근사적 95% 수준의 신뢰구간은

$$\left(5.5 - 1.96 \sqrt{5.5/4} \, , 5.5 + 1.96 \sqrt{5.5/4} \right) = (3.20, 7.80)$$

으로 산출됩니다. 이것은 앞에서 얻었던 신뢰구간인 $(3.45, 8.33)$과는 적지 않은 차이가 있습니다. 표본크기 n이 크지 않기 때문으로 생각됩니다.

근사적 신뢰구간의 유용성은 유한한 크기 n의 표본에서 '근사적' 신뢰구간들이 실제로 θ를 어느 정도 담아내느냐에 있습니다. 그 개념을 명확히 하기 위하여 일반적으로 신뢰구간의 행태를 다음과 같이 압축하여 나타냅니다.

포함확률(coverage probability) : 정의

신뢰구간 $(L(X_1, \cdots, X_n), U(X_1, \cdots, X_n))$이 실제 θ를 포함하는 확률

$$P\{\theta \in (L(X_1, \cdots, X_n), U(X_1, \cdots, X_n)) \mid \theta\}$$

를 포함확률이라고 합니다.

그러므로 근사적 수준 $1-\alpha$의 신뢰구간은 θ를 $1-\alpha$에 가까운 확률로 포함하여야 합니다. 포함확률을 평가해본 결과, 그것이 $1-\alpha$에 적지 않게 미달한다면 그것은 심각한 하자(瑕疵)입니다. 그러면 두번째 예인 포아송 표본($n = 4$)의 경우에서 θ에 관한 근사적 수준 95%의 신뢰구간

$$\left(\overline{X} - 1.96 \sqrt{\overline{X}/4} \, , \overline{X} + 1.96 \sqrt{\overline{X}/4} \right)$$

의 포함확률을 평가해보도록 합시다.

포함확률의 산출과정은 다음과 같습니다.

단계 1: $n(=4)$개의 변량 $x_1, \cdots, x_n$을 평균이 θ인 포아송 분포로부터 생성시킨다 (2.4절).

<표 3> 근사적 신뢰구간의 포함확률 평가를 위한 SAS/IML 프로그램

```
/* Coverage Evaluation for Poisson Parameter */
/* File Name:  conf3.iml                     */

proc iml;
   n = 4;  theta = 2;
   Nrepeat = 10000;
   Ncover=0;
   do repeat=1 to Nrepeat;
      sum = 0;
      do sample=1 to n;
         count = -1;
         sum1 = 0;
         do while (sum1 < 1);
            t = -(1/theta)*log(1-uniform(0));
            sum1 = sum1 + t;
            count = count + 1;
         end;
         X = count;
         sum = sum + X;
      end;
      Xbar = sum / n;
      U = Xbar + 1.96*sqrt(Xbar/n);
      L = Xbar - 1.96*sqrt(Xbar/n);
      cover=0;
      if theta <= U then do; if theta >= L then cover=1; end;
      Ncover = Ncover + cover;
      end;
      coverage = Ncover / Nrepeat;
      print theta Nrepeat coverage;
quit;
```

단계 2: 근사적 수준 95%의 신뢰구간 $\left(\overline{x} - 1.96\sqrt{\overline{x}/4}\,,\, \overline{x} + 1.96\sqrt{\overline{x}/4}\right)$ 을 계산한다.

단계 3: 단계 2의 구간에 파라미터 θ 가 포함되었는지의 여부를 검사한다.

단계 4: 단계 1·2·3을 N번 반복 시행하여 포함확률을 추정한다.

<표 3>은 포아송 파라미터 θ 가 2인 경우에서 근사적 수준 95% 신뢰구간의 포함확률을 $N(=10000)$ 번의 반복 시행을 통하여 산출하는 SAS/IML 프로그램입니다. 프로그램을 수행한 결과 포함확률이 89.2%로 추정되었습니다. 그러므로 신뢰구간이 95%의 신뢰수준을 표방하고는 있지만 실제로는 90% 남짓한 신뢰성만 갖고 있는 것이지요. 실망스러운 결과입니다. 다른 θ 값의 경우에 대하여도 포함확률을 추정한 결과는 다음과 같습니다($N = 10000$).

θ 값	1	2	3	4	5	6	7	8	9	10
포함확률	0.905	0.892	0.946	0.939	0.947	0.932	0.936	0.947	0.935	0.940

그러므로 $\theta \geq 3$ 인 경우에서는 "근사적" 95% 수준의 신뢰구간이 제대로의 포함확률을 확보하고 있는 것을 볼 수 있습니다. 그러나 $\theta < 3$ 인 경우에는 그렇지 않을 가능성이 다분히 있음을 경계해야 합니다. 좀 더 자세히 보기 위하여 $0.1 \leq \theta \leq 2.9$ 사이의 θ 에 대하여 0.1 간격으로 포함확률을 조사해볼 필요가 있습니다. 이것은 여러분에게 과제로 넘기겠습니다 (연습문제 7.6).

근사적 신뢰구간을 구하는 다른 방법으로 일반화 가능도 비를 이용하는 방법이 있습니다 (6.6절 참조). 그것의 내용은 k-파라미터 $\boldsymbol{\theta}$ 의 경우 일반화 가능도 비

$$\Lambda(x_1, \cdots, x_n; \boldsymbol{\theta}) = \frac{L(\hat{\boldsymbol{\theta}}\ ; x_1, \cdots, x_n)}{L(\boldsymbol{\theta}\ ; x_1, \cdots, x_n)}$$

의 2배 로그변환인 $2\log_e \Lambda(X_1, \cdots, X_n; \boldsymbol{\theta})$ 가 카이제곱 분포 $\chi^2(k)$ 를 *근사적으로* 따른다는 것입니다. 다시 말하면 $2\log_e \Lambda(X_1, \cdots, X_n; \boldsymbol{\theta})$ 가 근사적 추축량입니다. 따라서 $\boldsymbol{\theta}$ 에 관한 근사적 수준 $1-\alpha$ 의 신뢰구간은

$$\{\boldsymbol{\theta} \mid 2\log_e \Lambda(X_1, \cdots, X_n; \boldsymbol{\theta}) \leq \chi^2_{k,1-\alpha}\}$$

라고 할 수 있습니다. 구체적으로 $k=1$ 인 경우, θ 에 관한 근사적 수준 95%의 신뢰구간은 ($\chi^2_{1,0.95} = 3.8415$)

$$\left\{\theta \mid \frac{L(\hat{\theta}\ ; x_1, \cdots, x_n)}{L(\theta\ ; x_1, \cdots, x_n)} \leq e^{3.8415/2}\right\},$$

즉

$$\left\{\theta \mid L(\theta\ ; x_1, \cdots, x_n) \geq \frac{1}{6.8261} L(\hat{\theta}\ ; x_1, \cdots, x_n)\right\}$$

입니다. 여기서 $\frac{1}{6.8261}$ 은 대략 $\frac{1}{7}$ 로 볼 수 있겠습니다. 이것이 바로 4.1절에서 소개한 바 있는 7분의 1 규칙의 이론적 배경입니다. $k=2$ 인 경우, θ 에 관한 근사적 수준 95%의 신뢰구간은 ($\chi^2_{2,0.95} = 5.9915$), 같은 방법으로,

$$\left\{\boldsymbol{\theta} \mid L(\boldsymbol{\theta}\ ; x_1, \cdots, x_n) \geq \frac{1}{20.00} L(\hat{\boldsymbol{\theta}}\ ; x_1, \cdots, x_n)\right\}$$

이 되는데 이것이 바로 파라미터가 2개인 경우에 적용되는 20분의 1 규칙입니다.

$n=4$ 이고 $\overline{X} = 5.5$ 인 포아송 표본에서 θ 에 관한 근사적 95% 수준의 신뢰구간은 $(3.51, 8.13)$ 으로 계산됩니다 (4.2절 참조). 이것은 소표본에서 양측 가설검증을 뒤

집어 얻은 95% 수준의 신뢰구간 $(3.45, 8.33)$ 과 크게 다르지 않습니다.

새로운 예를 하나 들어보도록 하겠습니다. $x_1, \cdots, x_n$ 이 평균이 θ 이고 표준편차가 $0.1\,\theta$ 인 특수한 정규분포 $\mathrm{N}(\theta, (0.1\theta)^2)$ 으로부터의 관측된 임의표본이라고 합시다 (단,$\theta > 0$). 이 때 목표는 평균이면서 10배 표준편차인 θ 에 대한 신뢰구간을 구하는 것입니다. 로그 가능도가

$$l(\theta) = -\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{0.01\,\theta^2} - \frac{n}{2}\log_e(0.01\,\theta^2)$$

$$= -\frac{\sum_{i=1}^{n} x_i^2}{0.02\,\theta^2} + \frac{\sum_{i=1}^{n} x_i}{0.01\,\theta} - n\log_e\theta + constant$$

로 표현됩니다. 따라서 이 함수를 최대화하면 θ 에 대한 최대가능도추정치 $\hat{\theta}$ 를 얻을 수 있겠지요. 구체적으로 관측된 표본이

$$116.7,\ 94.7,\ 87.8,\ 110.4,\ 93.5\ \ (n=5)\ :\ \ \overline{x} = 100.62,\ \ s = 12.29$$

라고 합시다. θ 가 평균으로부터 100.62로 추정되고 표준편차로부터는 122.9로 추정되므로 이 두 값 사이에 보다 나은 추정치가 있으리라고 생각할 수 있습니다.

<그림 2>는 이 표본으로부터의 표준화 가능도를 그린 것이고 <표 4>는 최대가능도추정치를 산출하는 SAS/IML 프로그램입니다 (4.1절의 뉴튼-라프슨 방법 참조). 그 결과는 $\hat{\theta} = 100.81$ 로서 거의 표본 평균에 가깝습니다. 6.8261 분의 1 규칙에 의하여, 즉 일반화 가능도 비에 관한 대표본 이론에 의하여 θ 에 관한 근사적 95% 수준의 신뢰구간을 구한 결과 $(92.75, 110.40)$ 을 얻게 됩니다.

참고로, 모평균 θ 에 대한 통상적인 신뢰구간은 (7.1절 참조)

$$\overline{x} \pm t_{4,0.975}\frac{s}{\sqrt{n}} = 100.62 \pm 2.7765 \cdot \frac{12.29}{\sqrt{5}} = 100.62 \pm 15.26\,,$$

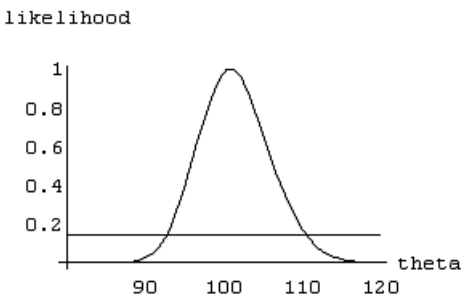
즉 $(85.36, 115.88)$ 입니다. 그러므로 일반화 가능도 비 이론에 따른 신뢰구간은 표본표준편차에 내포되어 있는 θ 에 관한 정보를 추가적으로 활용한다는 사실을 알 수 있습니다. 또 다른 참고로, 모분산 $0.01\,\theta^2$ 에 대한 통상적인 신뢰구간은 (7.1절 참조)

$$\left(\frac{(n-1)\,s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)\,s^2}{\chi^2_{n-1,\alpha/2}}\right) = \left(\frac{4 \cdot 12.29^2}{11.1}, \frac{4 \cdot 12.29^2}{0.484}\right) = (54.43, 1248.3)$$

입니다 (신뢰수준 95%). 이것을 θ 에 관한 구간으로 환산하면 $(73.8, 353.3)$ 입니다.

<표 4> 특수 정규표본에서 신뢰구간을 구하기 위한 SAS/IML 프로그램

```
/* Confidence interval from a special normal sample */
/* conf4.iml                                        */

proc iml;
   x = {116.7 94.7 87.8 110.4 93.5};
   s1 = sum(x);  s2 = ssq(x);   n = ncol(x);
   mean = s1/n;  sd = sqrt((s2 - s1*s1/n)/(n-1));
   print n mean[format=8.2] sd[format=8.2];
   theta = 90;
   maxtol = 0.0001;  maxiter = 10000;

   start ell;
      lik  = -s2/(2*0.01*theta**2)+s1/(0.01*theta)-n*log(theta);
      lik1 = s2/(0.01*theta**3)-s1/(0.01*theta**2)-n/theta;
      lik2 = -3*s2/(0.01*theta**4)+2*s1/(0.01*theta**3)+n/(theta**2);
   finish;

   iter = 0;
   tol  = 1;
   do while (iter <= maxiter & tol > maxtol);

      run ell;
      theta1 = theta - lik1/lik2;
      tol = abs(theta1 - theta);
      theta = theta1;
      iter = iter + 1;
   end;

   print iter theta[format=12.2] lik2[format=12.4] lik[format=12.4];
quit;
```



<그림 2> 특수 정규표본으로부터의 표준화 가능도와 신뢰구간

이제까지는 파라미터가 1개인 경우에서 근사적 신뢰구간을 다루었습니다만, 파라미터가 2개 이상인 경우에서의 근사적 신뢰구간도 같은 요령으로 개발될 수 있습니다. 일반화 가능도 비 검증 방법이 있으면 이에 대응하는 신뢰구간(또는 신뢰영역)이 존재하기 때문입니다. 그러나 여기서는 더 이상 다루지 않겠습니다.

7.4* 비편향 구간

이제까지는 θ에 관한 신뢰구간을 만드는 방법에 급급하여 어떤 구간이 좋은 구간인가에 대하여는 별로 언급하지 못하였습니다. 만약 두 개의 신뢰구간이 모두 θ에 관하여 동일한 포함확률을 가진다고 할 때 어느 구간이 더 좋다고 할 수 있을까요?

θ에 관한 임의구간 (L, U)는 θ를 포함하도록 의도되었기 때문에 $\theta'(\neq \theta)$는 가급적 포함하지 말아야 할 것입니다. 따라서 구간의 '비편향성(非偏向性)'을 다음과 같이 정의합니다.

비편향 구간(unbiased interval) : 정의

신뢰수준 $1-\alpha$ 의 임의구간 (L, U)가

$$P\{\theta' \in (L, U) \mid \theta\} \leq 1-\alpha, \text{ for } all \ \theta' \neq \theta$$

를 만족하면 비편향 구간이라고 합니다.

예를 들겠습니다. $X_1, \cdots, X_n$이 정규분포 $\mathrm{N}(\theta, \sigma_0^2)$으로부터의 임의표본인 경우($\sigma_0^2$은 미리 알려짐), θ에 관한 수준 $1-\alpha$ 의 신뢰구간인

$$\left(\overline{X} - z_{\alpha/2}\,\sigma_0 / \sqrt{n} \ , \ \overline{X} + z_{\alpha/2}\,\sigma_0 / \sqrt{n} \right) \tag{5}$$

이 비편향 구간인가를 검토해봅시다.

$$P\left\{ \overline{X} - z_{\alpha/2}\,\sigma_0 / \sqrt{n} \ \leq \ \theta' \ \leq \ \overline{X} + z_{\alpha/2}\,\sigma_0 / \sqrt{n} \mid \theta \right\} \tag{6}$$

$$= P\left\{ \frac{\overline{X} - \theta}{\sigma_0 / \sqrt{n}} - z_{\alpha/2} \ \leq \ \frac{\theta' - \theta}{\sigma_0 / \sqrt{n}} \ \leq \ \frac{\overline{X} - \theta}{\sigma_0 / \sqrt{n}} + z_{\alpha/2} \mid \theta \right\}$$

$$= \Phi\left(\frac{\theta' - \theta}{\sigma_0 / \sqrt{n}} + z_{\alpha/2} \right) - \left[1 - \Phi\left(- \frac{\theta' - \theta}{\sigma_0 / \sqrt{n}} + z_{\alpha/2} \right) \right]$$

이므로 (여기서 $\Phi(\cdot)$는 표준정규분포의 분포함수임), 이것을 θ'로 미분하면 0으로 놓으면 $\theta' = \theta$를 얻습니다. 즉,

$$\frac{\sqrt{n}}{\sigma_0}\left[\phi\left(\frac{\theta'-\theta}{\sigma_0/\sqrt{n}}+z_{\alpha/2}\right)-\phi\left(-\frac{\theta'-\theta}{\sigma_0/\sqrt{n}}+z_{\alpha/2}\right)\right]=0 \Rightarrow \theta'=\theta.$$

따라서 (6)은 $\theta'=\theta$ 일 때 최대값을 갖습니다. 그러므로 (5)는 비편향구간입니다.

그러나 괜찮은 신뢰구간 중에서도 비편향적이 아닌 것들이 다수 있습니다. 예를 들어 $X_1,\cdots,X_n$이 지수분포 Exponential(θ)로부터의 임의표본인 경우에서 θ에 관한 수준 $1-\alpha$의 신뢰구간인

$$\left(\frac{2n}{\chi^2_{2n,1-\alpha/2}}\cdot\overline{X}\,,\,\frac{2n}{\chi^2_{2n,\alpha/2}}\cdot\overline{X}\right)$$

은 비편향적이지 않습니다 (연습문제 7.7). 그러면 비편향구간 중에서 가장 좋은 것은 어떤 구간일까요? $\theta'(\neq\theta)$을 가급적 작은 확률로만 포함하는 구간일 것입니다.

균일최고정확 비편향구간(uniformly most accurate unbiased interval) : 정의

신뢰수준 $1-\alpha$를 갖는 임의의 비편향 임의구간 (L,U)에 대하여

$$P\{\theta'\in(L^*,U^*)\mid\theta\}\le P\{\theta'\in(L,U)\mid\theta\}\ \text{for all}\ \theta'\neq\theta$$

를 만족하는 신뢰수준 $1-\alpha$의 비편향구간 (L^*,U^*)을 균일최고정확 비편향 (UMA unbiased) 구간이라고 합니다.

다음 정리는 균일최고정확 비편향구간과 균일최강력 비편향(UMP unbiased) 검증 사이의 관계를 보여줍니다.

균일최고정확 비편향구간과 균일최강력 비편향검증의 관계 : 정리

$X_1,\cdots,X_n$이 확률(밀도)함수 $f(x;\theta)$로부터의 임의표본인 경우에서

$$\mathrm{H}_0:\theta=\theta_0 \quad \text{대} \quad \mathrm{H}_1:\theta\neq\theta_0$$

에 대한 유의수준 α의 균일최강력 비편향(UMP unbiased) 검증이 존재하고 그것의 채택역이 $L^*\le\theta_0\le U^*$라면, 구간 (L^*,U^*)는 균일최고정확 비편향 (UMA unbiased) 구간입니다.

이에 대한 증명은 생략하겠습니다. 천천히 따져보면 알 수 있습니다.

이 정리는 균일최강력 비편향검증을 뒤집어 신뢰구간을 얻어내는 것이 좋다는 것을 말해줍니다. 그러나 문제는 그런 검증이 존재하지 않거나 구현하기 어렵다는 것이지요.

7.5* 예측 구간

신뢰구간과 밀접한 관련이 있는 것이 예측구간입니다. 간단한 예측문제는 다음과 같습니다. $X_1, \cdots, X_n$ 이 어떤 확률(밀도)함수 $f(x;\theta)$ 로부터의 iid 관측변수라고 합시다. 그리고 X_{n+1} 을 동일한 분포로부터 미래 관측값을 나타내는 확률변수라고 합시다. 이 때, θ 값에 관계 없이 임의구간 $\left(L(X_1, \cdots, X_n), U(X_1, \cdots, X_n)\right)$ 으로 X_{n+1} 을 일정 수준이상의 확률로 담아낼 수 있다면 이 구간을 예측구간이라고 합니다. 즉,

예측구간 (predictive interval) : 정의

$$P\left\{ \left(L(X_1, \cdots, X_n) \leq X_{n+1} \leq U(X_1, \cdots, X_n)\right) \mid \theta \right\} \geq 1-\alpha, \quad \text{for } all\ \theta$$

를 만족하는 임의구간 $\left(L(X_1, \cdots, X_n), U(X_1, \cdots, X_n)\right)$ 을 수준 $1-\alpha$ 의 예측구간이라고 합니다.

신뢰구간과 마찬가지로 예측구간을 만드는 가장 핵심적인 방법은 추축량을 활용하는 것입니다. 가장 쉬운 예를 보겠습니다. $X_1, \cdots, X_n$, 그리고 X_{n+1} 을 평균이 θ 인 정규분포 $\mathrm{N}(\theta, \sigma_0^2)$ 으로부터 독립적으로 생성된다고 가정합시다 (σ_0^2 은 미리 알려짐). 그러면

$$\frac{X_{n+1} - \overline{X}}{\sigma_0 \sqrt{1+\frac{1}{n}}} \sim \mathrm{N}(0,1), \quad \text{for } all\ \theta$$

을 따릅니다. 따라서

$$P\left\{ -z_{\alpha/2} \leqq \frac{X_{n+1} - \overline{X}}{\sigma_0 \sqrt{1+\frac{1}{n}}} \leqq z_{\alpha/2} \mid \theta \right\} = 1-\alpha, \quad \text{for } all\ \theta$$

가 됩니다. 따라서

$$X_{n+1} \in \left(\overline{X} - z_{\alpha/2}\,\sigma_0 \sqrt{1+\frac{1}{n}}\ ,\ \overline{X} + z_{\alpha/2}\,\sigma_0 \sqrt{1+\frac{1}{n}} \right)$$

이라는 수준 $1-\alpha$ 의 예측구간을 얻습니다.

한편 σ_0^2 가 알려지지 않은 경우, X_{n+1} 에 관한 예측구간은

$$X_{n+1} \in \left(\overline{X} - t_{n-1,\alpha/2}\ s \sqrt{1+\frac{1}{n}}\ ,\ \overline{X} + t_{n-1,\alpha/2}\ s \sqrt{1+\frac{1}{n}} \right)$$

입니다. 여기서 s^2 은 $X_1, \cdots, X_n$ 으로부터 얻어지는 표본 분산입니다.

조금 어려운 문제로 포아송 분포의 경우를 보겠습니다. $X_1, \cdots, X_n$, X_{n+1}이 포아송 분포 Poisson(θ)로부터 독립적으로 생성된다고 합시다. 그러면

$$S_n = X_1 + \cdots + X_n \sim \text{Poisson}(n\theta)$$

이고, S_n은 X_{n+1}과 독립입니다. 그리고

$$S_{n+1} = S_n + X_{n+1} \sim \text{Poisson}((n+1)\theta)$$

입니다. 따라서

$$X_{n+1} \mid S_{n+1} = s_{n+1} \sim \text{B}\left(s_{n+1}, \frac{1}{1+n}\right)$$

을 따르게 됩니다. 왜냐하면

$$f_{X_{n+1} \mid S_{n+1}}(x_{n+1} \mid s_{n+1};\theta) = \frac{f_{X_{n+1}}(x_{n+1};\theta) \cdot f_{S_n}(s_n;\theta)}{f_{S_{n+1}}(s_{n+1};\theta)}$$

$$= \frac{\dfrac{e^{-\theta}\theta^{x_{n+1}}}{x_{n+1}!} \cdot \dfrac{e^{-n\theta}(n\theta)^{s_n}}{s_n!}}{\dfrac{e^{-(n+1)\theta}\{(n+1)\theta\}^{s_{n+1}}}{s_{n+1}!}} = \binom{s_{n+1}}{x_{n+1}} \left(\frac{1}{n+1}\right)^{x_{n+1}} \left(\frac{n}{n+1}\right)^{s_n}$$

이기 때문입니다. 따라서

$$\frac{\left(X_{n+1} - (S_n + X_{n+1})\dfrac{1}{1+n}\right)^2}{(S_n + X_{n+1})\dfrac{1}{1+n}\dfrac{n}{1+n}} \sim \chi^2(1), \text{ 근사적으로}$$

이라고 할 수 있습니다.

수치 예로서 $n = 4$, $X_1 = 3, X_2 = 10, X_3 = 5, X_4 = 4$ (즉, $S_4 = 22$) 인 경우에서 X_5에 대한 95% 수준의 예측구간은

$$\frac{\left(X_5 - (22 + X_5)\dfrac{1}{5}\right)^2}{(22 + X_5)\dfrac{1}{5}\dfrac{4}{5}} \leq 3.8415$$

로부터 얻을 수 있습니다. 즉 2차 방정식의 일반해로부터 $0.8187 \leq X_5 \leq 11.1417$이 나옵니다. X_5는 정수 값만을 취하므로 $1 \leq X_5 \leq 11$이라고 해야 하겠지요.

7.A 연습문제

7.1 $X_1, \cdots, X_n$을 지수분포 Exponential$(\theta, 1)$으로부터의 임의표본이라고 합시다. 즉

$$f_X(x;\theta) = e^{-(x-\theta)}, \quad x \geq \theta .$$

적절한 추축량을 제시하고 활용하여 θ에 관한 신뢰구간을 구해보세요.

7.2 $X_1, \cdots, X_n$을 균일분포 Uniform$(\theta - 0.5, \theta + 0.5)$으로부터의 임의표본이라고 합시다. 즉

$$f_X(x;\theta) = 1, \quad \theta - 0.5 \leq x \leq \theta + 0.5 .$$

적절한 추축량을 제시하고 활용하여 θ에 관한 신뢰구간을 구해보세요.

[힌트 : 충분통계량인 $X_{(1)}$과 $X_{(n)}$의 평균을 생각해보십시오.]

7.3 $H_0 : \theta = \theta_0$ 대 $H : \theta \neq \theta_0$ (단, $\theta \geq 1$)에 대한 검증에서, 통계량 $T \geq 0$에 의하여 유의수준 α의 기각역이

$$T \leq \theta_0 - \sqrt{\theta_0} \quad \text{or} \quad T \geq \theta_0 + \sqrt{\theta_0}$$

로 표현된다고 합시다. 관측된 T의 값이 5.5일 때, θ에 관한 수준 $1-\alpha$의 신뢰구간을 구하세요.

7.4 포아송 분포 Poisson(θ)로부터 임의표본 $3, 10, 5, 4$를 관측하였다고 합시다. 이 때, $\log_e \overline{X}$의 근사적 분포를 활용하여 $\log_e \theta \ (= \psi)$에 대한 95% 수준의 신뢰구간을 만드시오. 그리고, ψ에 관한 구간을 θ에 관한 구간으로 바꾸어 보세요.

7.5 크기 $n = 100$인 이항표본을 관측한 결과 성공비율이 $p = 0.57$이었다고 합시다. 이 때, $\log_e\{p/(1-p)\}$의 근사적 분포를 활용하여 $\log_e\{\theta/(1-\theta)\} \ (= \psi)$에 대한 95% 수준의 근사적 신뢰구간을 만드시오. 그리고, ψ에 관한 구간을 θ에 관한 구간으로 바꾸어 보세요.

7.6 포아송 파라미터 θ에 대한 수준 95%의 근사적 신뢰구간

$$\left(\overline{X} - 1.96\sqrt{\overline{X}/n} \, , \, \overline{X} + 1.96\sqrt{\overline{X}/n} \right)$$

의 실제 θ를 포함하는 확률을 $0.1 \leq \theta \leq 2.9$ 사이의 θ에 대하여 0.1 간격으로 계산해보세요. ① $n = 4$인 경우, ② $n = 16$인 경우.

7.7 $X_1, \cdots, X_n$이 지수분포 Exponential(θ)로부터의 임의표본인 경우에 θ에 관한 수준 $1-\alpha$의 신뢰구간인

$$\left(\frac{2n}{\chi^2_{2n,1-\alpha/2}} \cdot \overline{X} \, , \, \frac{2n}{\chi^2_{2n,\alpha/2}} \cdot \overline{X} \right)$$

가 비편향적이지 않음을 보이세요.

7.8 지수분포 Exponential(θ)로부터 크기 n의 임의표본 $X_1, \cdots, X_n$을 관측하였다 고 합시다. 동일한 분포로부터 독립적으로 X_{n+1}을 관측한다고 할 때, X_{n+1}에 대한 수준 95%의 예측구간을 만드시오.

[힌트 : $\dfrac{2X_{n+1}}{2\sum_{i=1}^{n} X_i}$의 분포를 잘 생각해보세요.]

탐구문제

이변량정규분포 BN($\mu_1, \mu_1, \sigma_1^2, \sigma_2^2, \rho$)로부터 크기$n\,(\geq 4)$의 임의표본을 관측하여, 모(母)상관계수 ρ에 관한 신뢰구간

$$h(\rho) \in (h(r) - 1.96/\sqrt{n-3}\,,\, h(r) + 1.96/\sqrt{n-3}\,)$$

으로부터 산출한다고 합시다. 여기서 r은 표본상관계수이고 피셔의 변환 $h(\rho)$는

$$h(\rho) = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho}, \quad -1 < \rho < 1$$

로 정의됩니다. ① 위 신뢰구간의 포함확률이 $\mu_1, \mu_1, \sigma_1^2, \sigma_2^2$과는 어떤 관련도 없음을 이론적으로 밝히세요. ② 몬테칼로 기법을 활용하여 위 신뢰구간의 포함확률을 구해보세요 ($n = 10, 20, 40$과 $\rho = 0.0, 0.5, 0.9$인 경우).

7.B 읽을만한 책

구간추정론에 대한 참고문헌은 가설검증론에 대한 참고문헌과 같습니다.

- Bickel, P.J. and Duksum, K.A. (1977) *Mathematical Statistics.*. Holden-Day. (Chapter 5)
- Casella, G. and Berger, R.L. (1990) *Statistical Inference*. Duxbury. (Chapter 9)
- Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. Chapman and Hall. (Chapter 7)