제 3 장 추정

§3.1 비율추정

3.1.1 서론

1장에서는 통계학의 기본적인 틀을 개괄적으로 언급했고, 2장에서는 통계학에 필요한 확률이론을 다루었다. 이제부터 정식으로 통계학을 논하겠는데, 3장에서 다룰 주제는 추정(estimation)이다.

추정 중에서는 비율 추정을 먼저 다루는데, 1장에 등장한 <사례 1.1>을 예제로 사용한다. 실제 득표율을 모집단의 비율 또는 줄여서 모비율(population proportion)이라 한다. 예를 들어, 김영삼 후보는 전체 N표의 (비고: N은 모집단의 크기) 42%인 .42N표를 얻었는데 이때 모비율은 0.42이다 (<그림 1.1, 1.2> 참조). 반면에, 추정치 (또는 예측치)로는 표본비율(sample proportion)을 사용했다. 표본의 크기 n은 약 2000인데, 이를 편의상 2000이라 하면 김영삼 후보는 2000표의 39.5%인 790표를 얻었으므로, 표본비율은 0.395이다.

<사례 1.1>을 예제로 사용하는 이유는 모비율이 알려진 특수한 사례이기 때문이다 (§1.1 참고). 따라서, 추정오차를 정확히 알 수 있다. 예를 들어 김영삼후보의 득표율에 대한 추정오차는 $0.395-0.42=-0.025$ 이다.

그러나, <사례 1.2>와 같은 일반적인 상황에서는 모비율이 알려지지 않는다. 부분의 비율로 전체의 비율을 추정하기 때문에 추정오차는 불가피한데, 모비율을 모르기 때문에 추정오차도 정확히 알 수는 없다. 다만, 추정오차의 확률분포로부터 추정치를 얼마나 신뢰할 수 있는가를 가름할 수 있을 뿐이다. 사실, <사례1.1>에서도 추정치 (또는 예측치)를 얻은 시점에서는 모비율이 알려져 있지 않았다. 이에 따라, 추정치를 발표할 때에 "95%의 신뢰수준에서 최대오차는 ± 0.022 (또는 $\pm 2.2\%$)"라고 오차의 범위도 함께 발표했다 (§1.2 참조). 이제, 추정치를 얻은 시점으로 돌아가서 (비고: 모비율이 알려지지 않은 시점임) 오차

의 범위에 대한 근거를 알아본다.

3.1.2 MLE $\hat{p}$

§1.6에서 MLE를 대표적인 추정방법으로 꼽았다. MLE는 한마디로 "관찰된 표본을 얻게 될 확률을 최대가 되게하는 모비율값"이다.

임의표본을 $\{Y_1,\cdots,Y_n\}$이라 하고 관찰된 표본을 $\{y_1,\cdots,y_n\}$이라 할 때 $P(Y_1=y_1,\cdots,Y_n=y_n)$을 LF(likelihood function)라 불렀다. 즉, LF는 모집단의 특정 부분집합인 $\{y_1,\cdots,y_n\}$이 표본으로 뽑힐 확률이다. $Y_1,\cdots,Y_n$은 모두 모분포를 따르는데, <그림 1.2>에서 편의상 $p_i=P(Y=i)$라 하자. 즉, 모비율을 $p_1,\cdots,p_5$로 표현한다. LF는 $p_1,\cdots,p_5$의 함수인데, LF가 최대가 되게 하는 $p_1,\cdots,p_5$ 값이 바로 MLE이다 (<비고 1.6.1> 참조).

이미 여러번 언급했듯이, 편의상 표본을 복원추출한 것으로 간주한다. 그러면, LF는 다음과 같이 다항분포를 따른다 (§2.1.6 참조).

$$L(p_1,\cdots,p_5)=Kp_1^{n_1}p_2^{n_2}p_3^{n_3}p_4^{n_4}p_5^{n_5} \tag{3.1.1}$$

$$K=n!/(n_1!n_2!n_3!n_4!n_5!) \tag{3.1.2}$$

위의 식에서 $n=\sum_{i=1}^{5} n_i$인데, $n_1,\cdots,n_5$는 후보별 득표수이다 (예 : $n_1=790$). 또한, $1=\sum_{i=1}^{5} p_i$이므로 식 (3.1.1)에서 p_5는 $(1-p_1-p_2-p_3-p_4)$와 동일하다.

식 (3.1.1)을 p_i에 대해서 ($i=1,2,3,4$) 편미분한 식을 0으로 놓고 풀면 $(p_i/n_i)=(p_5/n_5)$ 을 얻는다. 이로부터 (LF를 최대가 되게 하는 p_i값인) 최우추정치

$$\widehat{p_i}=\frac{n_i}{n} \tag{3.1.3}$$

를 얻는데 ($i=1,\cdots,5$), 이것이 바로 표본비율이다 (예 : $\widehat{p_1}=790/2000=0.395$). (비고: 최대값의 충분조건 확인과정은 생략함.)

또한, 식(3.1.3)의 n_i를 이에 대응하는 확률변수 N_i로 대체하면 최우추정량으로

$$\widehat{p}_i = \frac{N_i}{n} \tag{3.1.4}$$

를 얻는다 (비고: N_i는 식 (2.1.6)의 S_i와 동일).

<비고 3.1.1> 일반적으로 확률변수는 대문자로 표기하고 (예 : N_i) 대응하는 관찰치는 소문자로 표기하지만 (예 : n_i), 흔히 추정량과 추정치는 동일하게 표기한다 (예 : $\widehat{p}_i$).

<비고 3.1.2> 식 (3.1.1)에서 K를 누락시키더라도 동일한 MLE를 얻음 (식 (1.6.3) 참조). 즉, 복원추출시 추출하는 순서를 따지든 안 따지든 결과는 동일함 (§1.4 참조).

3.1.3 오차의 분포

추정오차를 $(\widehat{p}_i - p_i)$로 정의한다. 이에 따라, 오차가 양이면 overestimation을 음이면 underestimation을 의미한다. $\widehat{p}_i$가 최우추정치이면 오차는 상수이다 (예 : $\widehat{p}_1 - p_1 = 0.395 - 0.42 = -0.025$). 반면에 $\widehat{p}_i$가 최우추정량이면 오차는 확률변수가 되므로 확률분포를 거론하게 된다 (<비고 3.1.1> 참조).

식 (3.1.4)에서 $1 = \sum_{i=1}^{5} \widehat{p}_i$ (또는 $n = \sum_{i=1}^{5} N_i$) 이므로, $\widehat{p}_1, \cdots, \widehat{p}_5$는 독립이 아니다. 따라서 $\widehat{p}_1, \cdots, \widehat{p}_5$의 결합분포는 복잡하다. 그러나, 각각의 (주변 : marginal) 분포는 복잡하지 않다. §2.2.3에서 언급했듯이 N_i는 이항분포를 따르는데, 기대치는 np_i이고 분산은 np_iq_i이다 (단, $q_i = 1 - p_i$: §2.8.7 참조). 그런데, $n = 2000 \gg 1$이므로 중심극한정리를 활용하면 (§2.15.2 참조), $N_i \overset{A}{\sim} N(np_i, np_iq_i)$로 근사할 수 있다. 따라서,

$$\widehat{p}_i = \frac{N_i}{n} \overset{A}{\sim} N\left(p_i, \frac{p_iq_i}{n}\right) \tag{3.1.5}$$

$$\widehat{p}_i - p_i \overset{A}{\sim} N\left(0, \frac{p_iq_i}{n}\right) \tag{3.1.6}$$

$$\frac{\widehat{p}_i - p_i}{\sqrt{p_i q_i / n}} \overset{A}{\sim} N(0,1^2) \tag{3.1.7}$$

을 얻는다 (§2.9.1 참조).

3.1.4 오차의 범위

오차 ($\widehat{p}_i - p_i$)가 $\pm\varepsilon_i$이내에 들 확률이 0.95가 되게 하는 ε_i값을 구해보자.

$$\begin{aligned} 0.95 &= P(-\varepsilon_i < \widehat{p}_i - p_i < \varepsilon_i) \\ &= P\left(\frac{-\varepsilon_i}{\sqrt{p_i q_i / n}} < \frac{\widehat{p}_i - p_i}{\sqrt{p_i q_i / n}} < \frac{\varepsilon_i}{\sqrt{p_i q_i / n}}\right) \end{aligned} \tag{3.1.8}$$

이므로, 식 (3.1.7)과 표준 정규분포의 확률표로부터

$$1.96 \approx \varepsilon_i / \sqrt{p_i q_i / n} \tag{3.1.9}$$

를 얻는다.

식 (3.1.9)에 $n=2000$과 알려진 p_i값을 대입하면 아래의 결과를 얻는다 ($q_i = 1 - p_i$).

i	1	2	3	4	5
p_i	0.42	0.338	0.163	0.064	0.015
ε_i	0.0216	0.0207	0.0162	0.0107	0.0053

즉, p_i값이 0.5에 가까울수록 ε_i값이 커진다. 그런데, 추정치를 얻은 시점에서는 p_i값이 알려져 있지 않으므로, 식 (3.1.9)에 $p = q = 0.5$를 대입하여 ε의 최대값인 0.0219를 구한 것이 바로 "최대오차는 ±0.022"라고 언급한 것이다.

3.1.5 신뢰구간

"95% 신뢰수준에서 최대오차는 ±0.022"라는 표현 중에서 "최대오차는 ±0.022" 부분은 방금 설명했다. 이제 "95% 신뢰수준에서"부분을 설명한다.

신뢰구간(confidence interval)은 대표적인 구간 추정이다. 반면에, 앞에서 구한 MLE는 대표적인 점(point) 추정이다. 점 추정치인 최우추정치는 상수이고 점 추정량인 최우추정량

은 확률변수이듯이, 구간 추정치는 상수이고 구간추정량은 확률변수이다.

구간 추정량은 점 추정량으로부터 얻는다. 예를 들어, 신뢰수준(confidence level)이 95%인 구간 추정량은 다음과 같이 구한다. 식 (3.1.8)과 (3.1.9)로부터

$$0.95 \approx P(-1.96 < \frac{\widehat{p}_i - p_i}{\sqrt{p_i q_i / n}} < 1.96) \tag{3.1.10}$$

을 얻는데 (비고: "≈"를 사용한 이유는 중심극한정리에 따른 근사식이기 때문인데, 앞으로는 이를 무시하고 "="를 사용함), 우변의 괄호속을 정리하면

$$0.95 = P(\widehat{p}_i - 1.96\sqrt{p_i q_i / n} < p_i < \widehat{p}_i + 1.96\sqrt{p_i q_i / n}) \tag{3.1.11}$$

가 된다. 이때 유의할 점은 다음과 같다. 식 (3.1.10)에서는 부등식의 중간 항이 확률변수인 반면에, 식 (3.1.11)에서는 부등식의 첫 항과 끝 항이 확률변수이다 (비고: 중간항 p_i는 확률변수가 아님).

구간 추정량은 바로 식 (3.1.11)의 두 확률변수를 의미하는데, 이들을 하나로 묶어서

$$\widehat{p}_i \pm 1.96\sqrt{p_i q_i / n} \tag{3.1.12}$$

로 표현한다. 그리고, 구간 추정치는 식 (3.1.12)의 점 추정량 $\widehat{p}_i = N_i / n$을 점 추정치 $\widehat{p}_i = n_i / n$으로 대체하여 얻는다 (<비고 3.1.1> 참조).

그런데, 문제는 p_i를 몰라서 추정을 하고 있는 상황이므로, 식 (3.1.12)의 p_i와 q_i도 물론 모르는 값들이다. 따라서, 두 번째의 근사가 필요하다. (비고: 첫 번째의 근사는 중심극한정리를 사용한 것임.) 이때 점 추정치인 $\widehat{p}_i$로 p_i를 대체하는데, 이는 <비고 2.15.3>과 유사한 상황이다.

결과적으로, 95% 신뢰구간이라고 부르는 구간 추정치는 다음과 같다.

$$\frac{n_i}{n} \pm 1.96\sqrt{\frac{n_i}{n}\left(1 - \frac{n_i}{n}\right)/n} \tag{3.1.13}$$

예를 들어, p_1에 대한 95%신뢰구간은 0.395 ±0.0214 이고, p_5에 대한 95%신뢰구간은 0.012±0.0048 이다.

3.1.6 신뢰수준

95% 신뢰구간은 식 (3.1.11)로부터 얻었는데, 이때 95% (또는 0.95)는 엄연히 확률이

다. 그런데, 이를 확률이라 부르지 않고 신뢰수준이라고 부르는 이유는 다음과 같다.

식 (3.1.10), (3.1.11), (3.1.12)에서는 $\widehat{p}_i$가 확률변수(인 점 추정량)이므로 확률을 운운할 수 있다. 즉, $\widehat{p}_i \pm 1.96\sqrt{p_i q_i / n}$가 상수 p_i를 포함할 확률은 0.95이다. 다시 말해서, 확률변수 $\widehat{p}_i - 1.96\sqrt{p_i q_i / n}$은 p_i보다 작은 값을 가지고 또한 확률변수 $\widehat{p}_i + 1.96\sqrt{p_i q_i / n}$은 p_i보다 큰 값을 가질 (결합) 확률이 바로 0.95이다.

반면에, 식 (3.1.13)은 확률변수가 아니므로 확률을 운운할 수 없다. 예를 들어, p_1에 대한 95% 신뢰구간인 0.395 ± 0.0214가 p_1을 포함할 확률을 운운할 수는 없다. 만약, p_1이 확률변수라면 p_1이 0.395 ± 0.0214에 포함될 확률을 운운할 수 있을 것이다. 그러나, 추정 당시 $p_1 = 0.42$가 알려져 있지 않았을 뿐이지 p_1은 어디까지나 상수인 것이다.

<비고 3.1.3> 베이지안 통계학에서는 p_i를 확률변수로 취급함 (<비고 1.7.1> 참조).

그럼에도 불구하고, 구간추정치를 얼마나 “신뢰”할 수 있는가를 나타내기 위해서 0.95를 (확률수준이 아니라) “신뢰수준”이라고 부르는 것이다.

그런데, 신뢰구간 및 신뢰수준 등의 표현이 널리 쓰이다 보니까 오히려 일반인에게는 확률이라는 용어 (또는 개념)보다 더 친숙하게 느껴지게 되었다. 따라서, 불필요한 (또는 부적합한) 곳에 까지 신뢰수준이라는 표현을 사용하기도 하는데, 바로 “95% 신뢰수준에서 최대오차는 ± 0.022”라는 표현이 그 예이다. (비고: 식 (3.1.8)에서 0.95는 신뢰수준이 아니라 확률임.)

§3.2 정규 모분포 관련

3.2.1 MLE $\hat{\mu}$, $\widehat{\sigma^2}$

지금까지 LF를 $P(Y_1=y_1,\cdots,Y_n=y_n)$으로 정의했는데, 이는 $Y_1,\cdots,Y_n$이 이산 확률변수이었기 때문이다. 반면에 $Y_1,\cdots,Y_n$이 연속 확률변수인 경우에는 이들의 결합밀도함수를 LF로 정의한다 (§2.11.1 참조).

모분포가 정규분포라고 가정하면 <비고 2.7.1>에 의해서 $Y_1,\cdots,Y_n \sim iid\ N(\mu,\sigma^2)$이다 (§2.5.12 참조). 따라서, LF는 다음과 같다 (식 (2.11.2), (2.5.1) 참조).

$$L(\mu,\sigma^2)=\prod_{i=1}^{n} f(y_i)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2} \tag{3.2.1}$$

편의상, 식 (3.2.1)에 자연대수(natural log)를 취하면

$$\ln L(\mu,\sigma^2)=K-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \tag{3.2.2}$$

가 되는데, K는 미분할 때 0이 되는 항이다. (비고: 표본추출의 순서까지 따져서 식 (3.2.1)에 $n!$을 곱하면 이 또한 K에 포함됨.) 대수함수는 1:1 함수이므로 $\ln L(\mu,\sigma^2)$을 최대가 되게 하는 μ와 σ^2의 값은 $L(\mu,\sigma^2)$도 최대가 되게 한다. 식 (3.2.2)를 μ와 σ^2에 대해서 편미분한 식을 0으로 놓고 연립으로 풀면 최우추정치로 $\hat{\mu}=\sum_{i=1}^{n} y_i/n \equiv \overline{y}$와 $\widehat{\sigma^2}=\sum_{i=1}^{n}(y_i-\overline{y})^2/n$을 얻는다. (비고: 최대값의 충분조건 확인과정은 생략함.) 그리고, y_i와 $\overline{y}$를 각각 Y_i와 $\overline{Y}$로 대체하면 아래의 최우추정량을 얻는다 (<비고 3.1.1> 참조).

$$\hat{\mu}=\overline{Y} \left|_{K}^{\hat{H}}=\frac{\sum_{i=1}^{n} Y_i}{n}\right|_{\tilde{N}}^{H} \tag{3.2.3}$$

$$\widehat{\sigma^2}=\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n} \tag{3.2.4}$$

다소 비현실적이기는 하지만, 만약 σ^2이 알려져 있으면 식 (3.2.2)는 μ에 대해서만 미분하면 되는데, 결과는 (3.2.3)과 같다. 반면에 μ가 알려져 있는 경우에는 식 (3.2.2)를 σ^2에

대해서 미분해서 $\widehat{\sigma^2} = \sum_{i=1}^{n}(y_i - \mu)^2/n$ 을 얻는데, 이에 대응하는 최우추정량

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{n} \tag{3.2.5}$$

은 식 (2.15.8)과 일치한다.

3.2.2 $\widehat{\sigma^2}$과 S^2

이제 근본적인 문제를 짚고 넘어갈 때가 되었다. 예를 들어, 모분산 σ^2에 대한 점 추정량으로 식 (3.2.4)의 $\widehat{\sigma^2}$과 식 (2.15.9)의 S^2중에서 어느 것을 선택할 것인가 하는 문제이다.

사실 이 문제에 대한 딱부러진 해답은 없는데, 그 이유는 이 문제가 결국 다기준(multi-criteria)의사결정 문제이기 때문이다. 즉, 서로 상충되는 기준들간의 절충이 필요한데, 이때 절충방법은 각자의 선호도에 따라 다를 수 있기 때문이다.

결과부터 언급하면 다음과 같다. (자세한 내용은 §3.3에서 다룸.)

$$E[(\widehat{\sigma^2} - \sigma^2)^2] < E[(S^2 - \sigma^2)^2]$$
$$E(\widehat{\sigma^2} - \sigma^2) \neq 0, \quad E(S^2 - \sigma^2) = 0$$

즉, 첫 번째 기준인 $\min E[(\text{추정오차})^2]$에 의하면 $\widehat{\sigma^2}$이 낫지만, 두 번째 기준인 $\min|E(\text{오차})|$에 의하면 S^2이 낫다.

§3.3 점추정량의 선택

먼저 용어를 정의한다. 모집단의 특성치인 모비율(p_i), 모평균(μ), 모분산(σ^2) 등을 모수(population parameter)라 하는데, 이들은 물론 모분포를 표현할 때 사용되는 parameter 이기도 하다 (<비고 1.7.1> 참조).

편의상, 모수를 θ로 표현하고, θ에 대한 점 추정량을 $\hat{\theta}$이라 하자. 한마디로 추정오차는 0에 가까울수록 좋다. 그런데, 이 책에서 오차로 정의한 (§3.1.1 참조) $\square\ \hat{\theta}-\theta\ \square$는 확률변수이므로 오차가 0에 얼마나 가까운지는 확률적으로 (또는, 기대치적으로) 운운할 수 밖에 없다.

첫째로, 표본의 크기 n이 클수록 $E[(\hat{\theta}-\theta)^2]$또는 $E(|\hat{\theta}-\theta|)$가 점점 작아지다가, 극단적으로 $n\to\infty$이면 0이 되는 경우에 $\hat{\theta}$을 일치(consistent)추정량이라 하는데, 사실상 이 책에 등장하는 모든 추정량이 이에 해당된다.

둘째로, 같은 크기의 표본을 가지고도 $E[(\hat{\theta}-\theta)^2]$또는 $E(|\hat{\theta}-\theta|)$ 중에서 어느 것을 (또는, 제 3의 것을) 기준으로 사용할 것인가 하는 문제가 발생한다. 이는 마치 평균과 중앙값 중에서 어느 것을 모집단의 대표값으로 사용할 것인가 하는 상황과 같다 (§2.8.2, §2.8.3 참조). 모평균 μ는 $E[(Y-y_0)^2]$을 최소가 되게 하는 y_0값이고, 중앙값 m은 $E(|Y-y_0|)$을 최소가 되게 하는 y_0값이므로, μ와 m이 모두 나름대로 의미가 있기는 하지만 이미 우리는 은연중에 μ를 선호하고 있다. 특히, 모분포가 정규분포라고 가정하는 상황에서는 정규분포의 parameter인 μ(와 σ^2)의 사용이 당연시 되고 있다 (<비고 2.8.1> 참조). 이와 유사하게, 관행상 $E[(Y-y_0)^2]$에 대응하는 $E[(\hat{\theta}-\theta)^2]$을 기준으로 사용하는데, 이를 MSE(mean square error)라 부른다.

셋째로, MSE는 다음과 같이 표현할 수 있다 (식 (2.8.3) 참조).

$$E[(\hat{\theta}-\theta)^2]=V(\hat{\theta}-\theta)+\{E(\hat{\theta}-\theta)\}^2 \tag{3.3.1}$$

즉, $MSE=V(\text{오차})+E(\text{오차})^2$인데, 이때 $E(\text{오차})$를 편의(bias)라 부른다. 그리고, $E(\text{오차})=0$인 경우에 $\hat{\theta}$를 불편(unbiased)추정량이라 부른다. 즉, 불편추정량 $\hat{\theta}$의 정의는 다음과 같다.

$$E(\hat{\theta})=\theta \tag{3.3.2}$$

넷째로, $E(\hat{\theta}_1)=E(\hat{\theta}_2)=\theta$일 때 $V(\hat{\theta}_1)<V(\hat{\theta}_2)$이면, $\hat{\theta}_1$이 ($\hat{\theta}_2$에 비해서) 상대적으

로 효율적(efficient)이라고 한다. (비고: $V(\widehat{\theta}_i-\theta)=V(\widehat{\theta}_j)$, <비고 2.9.1> 참조.) 그리고, θ에 대한 모든 불편추정량들 중에서 분산이 최소인 것을 MVUE(minimum variance unbiased estimator)라 부른다. 즉, MVUE는 불편추정량이라는 제약(constraint) 하에 MSE를 최소화(minimize)하는 추정량이다. 반면에, 아무런 제약없이 MSE를 최소화하는 추정량을 min MSE 추정량이라 한다.

투자 대안들(alternatives) 중에서 하나를 선택하는 문제와 비교해 보자. 투자의 위험도(risk)는 따지지 않고 무조건 기대 수익률이 최대인 대안을 선택하는 것은 min MSE 방법과 유사하다. 반면에 정기예금 및 국채와 같이 위험도가 0에 가까운 대안들 중에서 기대 수익률이 최대인 대안을 선택하는 것은 MVUE 방법과 유사하다. 보수적인 의사결정이 보편적으로 선호되듯이, min MSE 방법에 비해서 MVUE 방법이 선호된다.

MVUE를 구하는 방법은 §3.4에서 다룬다. 결과부터 언급하면, 식 (3.1.4)의 $\widehat{p}_i=N_i/n$, 식 (3.2.3)의 $\hat{\mu}=\overline{Y}$, 식 (2.15.8)과 (2.15.9)의 S^2은 모두 MVUE 이다. 즉, 이미 MLE 방법으로 구한 추정량들 중에서 식 (3.2.4)의 $\widehat{\sigma^2}$을 제외한 모든 것이 (결과론적으로) MVUE이다.

그렇다면, MLE는 무엇인가? 최우추정치는 (θ의 함수인) LF를 최소화하는 θ값이고, 최우추정량은 최우추정치의 y_i를 Y_i로 대체한 것이다. 그러니까, 의사결정의 기준 자체가 다르다. 즉, □ min MSE□를 기준으로 사용하는 것이 아니라 □max LF□를 기준으로 사용하는 것이다. MLE 방법을 투자 문제에 비유한다면 □max LF□라는 기준은 나름대로 설득력이 있는 투자지침에 해당 된다. 그리고, 이 투자지침을 따르면 p와 μ에 대해서는 MVUE 방법과 동일한 대안을 선택하게 되지만, σ^2에 대해서는 약간 다른 대안을 선택하게 된다.

아래의 표는 (μ가 알려지지 않은) 정규 모분포의 σ^2에 대한 추정량 3개를 비교한 것이다. (비고: 편의상 $\sum_{i=1}^{n}(Y_i-\overline{Y})^2$을 SS (sum of squares)라 부르면, 식 (2.15.11)와 <비고 2.8.10>에 의해서 $E(SS)=(n-1)\sigma^2$, $V(SS)=2(n-1)\sigma^4$.)

추정량	$\frac{SS}{n-1}$		$\frac{SS}{n}$		$\frac{SS}{n+1}$
E(오차)^2	0	<	$\frac{1}{n^2}\sigma^4$	<	$\frac{4}{(n+1)^2}\sigma^4$
V(오차)	$\frac{2}{n-1}\sigma^4$	>	$\frac{2(n-1)}{n^2}\sigma^4$	>	$\frac{2(n-1)}{(n+1)^2}\sigma^4$
MSE	$\frac{2}{n-1}\sigma^4$	>	$\frac{2n-1}{n^2}\sigma^4$	>	$\frac{2}{n+1}\sigma^4$

MVUE인 $SS/(n-1)$은 E(오차)$=0$ 이지만 MSE는 제일 크다. 반면에, MLE인 SS/n의 MSE는 MVUE보다 작지만 E(오차)$\neq 0$이다. 그리고, $SS/(n+1)$의 MSE는 MLE보다도 작지만 $|E$(오차)$|$는 더 커진다. (비고: $SS/(n+1)$은 SS에 n의 함수를 곱한 형태의 추정량 중에서 MSE가 최소인 것임.)

§3.4 MVUE 구하는 방법

MVUE는 최소충분(minimal sufficient)통계량의 함수라고 알려져 있다. 그리고, 최소충분통계량을 구하는 방법도 알려져 있다. 그러나, 함수형태는 주먹구구식으로 찾을 수 밖에 없다. 즉, 최소충분통계량의 함수 중에서 불편추정량이 되는 것을 자동적으로 찾아 주는 방법은 없다.

최소충분통계량이란 모수 θ를 추정하는데 필요한 최소한의 표본정보를 의미하는데 (<비고 1.3.1> 참조), 지금까지 등장한 예는 다음과 같다. §1.6 의 <사례 1.3>에서는 □임의 표본에 속한 꼬리표를 단 동물의 수□이고, §3.1 의 비율 추정에서는 식 (3.1.4)의 □ N_i□이며, §3.2 의 모평균 추정에서는 식 (3.2.3)의 □ $\sum_{i=1}^{n} Y_i$□이고 모분산 추정에서는 식 (2.15.9)의 SS 속에 들어 있는 □ $\sum_{i=1}^{n} Y_i$ 와 $\sum_{i=1}^{n} {Y_i}^2$□이다.

<비고 3.4.1> $SS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} {Y_i}^2 - (\sum_{i=1}^{n} Y_i)^2/n$

최소충분통계량의 의미를 표본 $\{Y_1, Y_2, Y_3\}$로 설명한다 (§2.15.4 참조).

$S_i = {Y_1}^i + {Y_2}^i + {Y_3}^i$라 하면, $\{Y_1, Y_2, Y_3\}$를 $\{S_1, S_2, S_3\}$로 대체하더라도 표본에 담긴 정보의 손실은 없다. 그런데, μ를 추정하기 위한 정보는 $\{S_1\}$만으로 충분하다. 즉, 일단 S_1이 있으면 S_2와 S_3는 불필요하다. 또한, σ^2을 추정하기 위한 정보는 $\{S_1, S_2\}$만으로 충분하다.

MVUE가 최소충분통계량의 함수로 알려져 있듯이, MLE는 충분통계량의 함수로 알려져 있다. 충분통계량에는 꼭 필요한 최소한의 정보 외에 불필요한 정보도 포함될 수 있다. (즉, 최소충분통계량은 충분통계량의 일종의 부분집합이다.) 그러나, 이는 어디까지나 일반적인 경우일 뿐이고, 지금까지 등장한 MLE 모두 최소충분통계량의 함수이다. 이러한 이유로 MLE가 불편추정량이면 그 자체가 MVUE가 되기도 하고 (예: $\widehat{p}_i$ 와 $\hat{\mu}$), MLE가 불편추정량이 아닌 경우에는 이를 불편추정량이 되도록 손질하여 MVUE를 얻을 수도 있다 (예: SS/n의 분모 n을 $n-1$로 대체함).

마지막으로, 최소충분통계량을 구하는 방법을 정규 모분포를 예로 들어서 설명한다. 지금까지 임의표본을 $\{Y_1, \cdots, Y_n\}$이라 하고 관찰된 표본을 $\{y_1, \cdots, y_n\}$이라 했는데, 이제 또 다른 관찰된 표본을 $\{y_1', \cdots, y_n'\}$이라 하자. 식 (3.2.1)에 의해서 각각의 LF는

$$L=\prod_{i=1}^{n} f(y_i)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

$$L'=\prod_{i=1}^{n} f(y_i')=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i'-\mu)^2}$$

이므로, 이들의 비율은 다음과 같다.

$$\frac{L}{L'}=e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}y_i^2-\sum y_i'^2\right)}e^{-\frac{\mu}{\sigma^2}\left(\sum_{i=1}^{n}y_i-\sum y_i'\right)} \tag{3.3.3}$$

먼저, μ의 값을 변화시키더라도 $\frac{L}{L'}$의 값이 변화하지 않는 필요충분조건을 구하면 $\sum_{i=1}^{n} y_i=\sum_{i=1}^{n} y_i'$을 얻는데, 이를 확률변수로 표현한 $\square\sum_{i=1}^{n} Y_i\square$가 바로 μ에 대한 최소충분통계량이다. 다음, σ^2의 값을 변화시키더라도 $\frac{L}{L'}$의 값이 변화하지 않는 필요충분조건을 구하면 $\sum_{i=1}^{n} y_i^2=\sum_{i=1}^{n} y_i'^2$과 $\sum_{i=1}^{n} y_i=\sum_{i=1}^{n} y_i'$을 얻는데, 이들을 확률변수로 표현한 $\square\sum_{i=1}^{n} Y_i^2$ 과 $\sum_{i=1}^{n} Y_i\ \square$가 바로 σ^2에 대한 최소충분통계량이다.

§3.5 MLE 의 속성

3.5.1 MLE 와 MVUE

§3.2까지는 MLE를 대표적인 추정방법이라 했는데, §3.3에서 MVUE가 등장하면서부터는 MLE가 밀리는 것같은 인상을 준다. 그럼에도 불구하고 MLE는 여전히 가장 중요한 방법인데 그 이유는 다음과 같다.

첫째로, MLE는 구하기가 쉽다. 즉, LF 만 있으면 이를 미분해서 MLE를 구할 수 있다.

둘째로, θ의 MLE $\hat{\theta}$만 구하면 θ의 함수 $g(\theta)$의 MLE는 자동적으로 $g(\hat{\theta})$가 되는데 (단, $g(\cdot)$은 1:1함수), 이를 MLE의 불변성(invariance) 속성이라 한다. 예를 들어, σ^2의 MLE가 SS/n이면 σ의 MLE는 $\sqrt{SS/n}$이다.

셋째로, MVUE는 최소충분통계량의 함수라고 했는데, 함수형태에 대한 힌트를 MLE의 형태로부터 얻을 수 있다. 특히, §3.4에서 보았듯이 MLE가 최소충분통계량의 함수인 경우에는 MLE를 불편추정량이 되도록 간단히 손질만 하면 바로 MVUE를 얻는다. 그러니까, MVUE를 원하는 경우에조차 MLE가 도움이 된다.

넷째로, 중심극한정리를 확장하면 모든 MLE에 적용된다 (§2.15.2 참조). 즉, 모든 최우추정량은 점근적으로(asymptotically) 정규분포를 따른다. 더우기, 최우추정량은 점근적으로 불편추정량일뿐더러 분산은 점근적으로 이론적인 하한치와 일치한다 (구체적인 내용은 §3.5.2 참조).

이 속성은 사실 대학원 수준에 가서야 제대로 진가를 발휘하는데, 그 이유는 다음과 같다. 학부 수준인 이 책에 지금까지 등장한 모수인 p_i, μ, σ^2에 대해서는 MVUE 뿐만 아니라 (MVUE의) 확률분포조차 비교적 쉽게 구할 수 있었다. 그러나, 문제가 복잡해지면 MVUE를 구하기 어려운 경우가 발생하기도 하고 또한 MVUE는 구하더라도 그 확률분포를 구하기 어려운 경우가 발생한다. (비고: 확률분포를 알아야 추정치에 대한 신뢰수준을 운운할 수 있음.) 그러나, MLE를 구하기만 하면 정규분포를 근사분포로 사용할 수 있는데, 이는 MLE의 확률분포를 구하기 어려운 경우 뿐만 아니라 확률분포를 구하더라도 그 분포가 사용하기에 복잡한 경우에도 해당된다. (이는 예를 들어 이항분포를 정규분포로 근사하는 것과 유사하다: §2.15.2 참조.)

나아가서, LF로부터 MLE를 얻는 과정이 분석적으로(analytically) 어려운 경우조차 발생하는데, 이때에도 수치적으로(numerically) 최우추정치를 얻을 수 있을 뿐만 아니라, 최우추정치에 대한 신뢰수준도 거론할 수 있다. 신뢰수준을 거론할 때 (또는 신뢰구간을 구할

때) 사용되는 것은 최우추정량의 점근적 분산인데, 놀라운 사실은 이 점근적 분산이 단순한 근사치가 아니라 이론적으로 밝혀진 하한치라는 점이다.

3.5.2 MLE의 점근 분포

모수 θ의 최우추정량인 $\hat{\theta}$의 점근(asymptotic) 분포는 다음과 같다.

$$\hat{\theta} \overset{A}{\sim} N(\theta,\ I(\theta)) \tag{3.5.1}$$

즉, $\hat{\theta}$의 분포는 n이 클수록 점점 정규분포에 가까와지는데, 평균은 θ이고 (따라서, $\hat{\theta}$은 점근적으로 불편추정량이고), 분산은 관례상 $I(\theta)$로 표기한다.

$I(\theta)$는 사실 MVUE 와 관련이 있다. 모수 θ의 MVUE를 $\tilde{\theta}$라 하면, $V(\tilde{\theta})$는 θ에 대한 모든 불편추정량들 중에서 최소이다. 이때

$$V(\tilde{\theta}) \geq I(\theta) \tag{3.5.2}$$

가 성립하는데, 이를 Cramer-Rao 부등식이라 한다. 그러니까, $\tilde{\theta}$가 MVUE 더라도 □ $V(\tilde{\theta}) > I(\theta)$ □가 가능하다. 반면에, 불편추정량 $\tilde{\theta}$가 MVUE 이기 위한 충분조건은 □ $V(\tilde{\theta}) = I(\theta)$ □이다.

<비교 3.5.1> §3.3에서 $E(\tilde{\theta}_1) = E(\tilde{\theta}_2) = \theta$ 이고 $V(\tilde{\theta}_1) < V(\tilde{\theta}_2)$일 때 $\tilde{\theta}_1$이 $\tilde{\theta}_2$ 에 비해서 □상대적으로□ 효율적이라고 했다. 그런데, 만약 $E(\tilde{\theta}_3) = \theta$ 이고 $V(\tilde{\theta}_3) = I(\theta)$ 이면, $\tilde{\theta}_3$를 (절대적으로) 효율적이라 한다. 이에 따라, 모든 MLE를 점근적으로 효율적(efficient)이라 일컫는 다.

$I(\theta)$의 정의는 다음과 같다.

$$I(\theta) = \frac{-1}{n \cdot E\left[\frac{\partial^2 \ln f(Y)}{\partial \theta^2}\right]} \tag{3.5.3}$$

$f(Y)$는 연속 모분포의 밀도함수 $f(y)$에서 y를 Y로 대체한 것이다. 예를 들어, 정규 모분포의 경우에는

$$\ln f(Y) = K - \frac{1}{2}\ln\sigma^2 - \frac{(Y-\mu)^2}{2\sigma^2} \qquad (3.5.4)$$

인데, K는 미분할 때 0이 되는 항이다 (식 (3.2.2) 참조). $\Theta = \mu$ 인 경우에는 $\frac{\partial^2}{\partial\mu^2}\ln f(Y) = \frac{-1}{\sigma^2}$ 이므로 $I(\mu) = \sigma^2/n$ 을 얻는데, 이는 $V(\overline{Y})$ 와 동일하다. 반면에, $\Theta = \sigma^2$ 인 경우에는

$$\frac{\partial^2}{\partial(\sigma^2)^2}\ln f(Y) = \frac{1}{2\sigma^4} - \frac{(Y-\mu)^2}{\sigma^6}$$

이므로 $E[(Y-\mu)^2] = \sigma^2$ 임을 활용해서, $I(\sigma^2) = 2\sigma^4/n$ 을 얻는다. (비고: $V(SS/(n-1)) = 2\sigma^4/(n-1) > I(\sigma^2)$.)

<비고 3.5.2> 식 (3.5.3)의 우변에서 2차 편미분 □$\partial^2 \ln f(Y)/\partial\Theta^2$□이 존재하지 않으면 대신 □$-[\partial \ln f(Y)/\partial\Theta]^2$□를 사용할 수 있음.

또한, 식 (3.5.3)은 다음과 같이 확장된다. 모수 또는 모분포의 parameter가 두 개 이상일 때 이들의 최우추정량들 간의 점근적 공분산을 유사한 방법으로 구할 수 있다. 예를 들어, 정규 모분포의 경우에 $\hat{\mu}$ 와 $\widehat{\sigma^2}$간의 공분산은 점근적으로 다음과 같다.

$$I(\mu, \sigma^2) = \frac{-1}{n \cdot E[\frac{\partial^2 \ln f(Y)}{\partial\mu\,\partial(\sigma^2)}]} = 0$$

즉, <비고 2.15.2>에 의해서 $\overline{Y}$와 SS가 서로 독립이므로 $\hat{\mu} = \overline{Y}$ 와 $\widehat{\sigma^2} = SS/n$도 서로 독립이다. 따라서 $\hat{\mu}$ 와 $\widehat{\sigma^2}$간의 공분산은 0인데 이는 물론 $n \to \infty$ 일 때에도 유효하다.

마지막으로, $I(\Theta)$에서 I는 Information을 의미하는데, 이는 $I(\Theta)$가 추정오차에 대한 정보를 제공한다는 뜻이다. 또한, 공분산까지 포함시킨 행렬을 Information matrix 라 하는데, 예를 들면 $\begin{bmatrix} I(\mu) & I(\mu,\sigma^2) \\ I(\sigma^2,\mu) & I(\sigma^2) \end{bmatrix}$ 이다.

§3.6 Moment 방법

MVUE와 MLE를 모두 구할 수 없을 때 시도해볼 만한 방법이 MMM(the method of matching moments) 인데, 이는 한마디로 모집단의 k^{th}moment 인 $E(Y^k)$와 표본의 k^{th} moment 인 $\sum_{i=1}^{n} y_i{}^k/n$을 같다고 놓은 식을 푸는 것이다. (단, $k=1,2,\cdots$ 의 순서로 식을 세우되 필요한 개수만 사용함.) 예를 들면

$$\mu = E(Y) = \sum_{i=1}^{n} y_i/n = \overline{y}$$

$$\sigma^2 + \mu^2 = E(Y^2) = \frac{\sum_{i=1}^{n} y_i{}^2}{n} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n} + (\overline{y})^2$$

으로부터 moment 추정치인 $\hat{\mu} = \overline{y}$와 $\widehat{\sigma^2} = \sum_{i=1}^{n} (y_i - \overline{y})^2/n$ 을 얻은 다음, y_i를 Y_i로 대체해서 moment 추정량인 $\hat{\mu} = \overline{y}$와 $\widehat{\sigma^2} = SS/n$ 을 얻는다.

moment 추정량에 대해서 유일하게 알려진 속성은 일치추정량이라는 것인데 (§3.3 참조), 물론 MVUE와 MLE도 일치추정량이다.

한가지 유의할 점은 다음과 같다. 위에서 moment 추정량 $\hat{\mu} = \overline{y}$와 $\widehat{\sigma^2} = SS/n$ 을 얻는 과정에서 모분포에 대한 아무런 가정을 하지 않았다. 그런데도 결과는 정규 모분포라는 가정하에 얻은 최우추정량과 일치한다. 이는 다음과 같은 가능성을 암시한다. 정규 모분포를 가정하고 얻은 최우추정량들은 정규 모분포라는 가정이 다소 무리가 있더라도 별로 영향을 받지 않는다는 점인데, 이를 MLE의 robustness 속성이라 한다. (즉, MLE는 모분포가 달라지더라도 이에 별로 민감하지 않다는 뜻이다.)

<비고 3.6.1> OR(operations research)의 용어를 빌리면, MVUE와 MLE는 최적해인 반면에 moment 추정량은 heuristic 해이다. 즉, MVUE는 불편추정량이 라는 제약하에 MSE를 최소화시키는 최적해이고, MLE는 LF를 최대화 시키는 최적해인 반면에, MMM은 단지 일리가 있는 heuristic 방법이라 고 할 수 밖에 없다.

§3.7 신뢰구간

3.7.1 모집단 하나의 경우

이 책에는 두 종류의 구간추정이 등장하는데 첫째는 이미 §3.1.5에 등장했던 신뢰구간이고 둘째는 §6.4.7에 등장할 예측구간(prediction interval)이다.

먼저, PQ(pivotal quantity)를 정의한다. 예를 들어, 식 (3.1.10) 우변의 부등식에서 축(pivot)의 역할을 하고 있는 중간 항이 바로 PQ이다. 일반적으로, 모수 θ에 대한 신뢰구간을 구할 때 사용하는 PQ는 확률변수인데 그 분포가 (정확하게 또는 근사적으로) 알려져 있어야 되고, 또한 θ의 함수이어야 된다.

정규 모분포의 모평균 μ에 대한 95% 신뢰구간을 구해보자. PQ로 $Z \equiv (\overline{Y} - \mu)/(\sigma/\sqrt{n})$를 사용하면, $Z \sim N(0,1^2)$이므로

$$0.95 = P(-1.96 < \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} < 1.96)$$

$$= P(\overline{Y} - 1.96\,\sigma/\sqrt{n} < \mu < \overline{Y} + 1.96\,\sigma/\sqrt{n})$$

을 얻는데, $\square\ \overline{Y} \pm 1.96\sigma/\sqrt{n}\ \square$이 바로 μ에 대한 구간 추정량이다. 물론, $\overline{Y}$를 $\overline{y}$로 대체하면 구간 추정치인 95% 신뢰구간을 얻는다. 그러나 이는 모분산 σ^2이 알려진 경우에나 사용할 수 있다. σ^2을 모르는 경우에는 관행상 이를 MVUE인 $S^2 = SS/(n-1)$로 대체하는데 (§2.15.5 참조), 이에 따라 식 (2.15.13)의 $T_{n-1} = (\overline{Y} - \mu)/(S/\sqrt{n})$을 PQ로 사용하게 된다. 예를 들어, $n = 10$인 경우에는 t분포의 확률분포로부터 $0.025 = P(T_9 > 2.262) = P(T_9 < -2.262)$를 얻으므로, 결국 구간 추정량은 $\square\ \overline{Y} \pm 2.262\,S/\sqrt{n}\ \square$이 된다. 그리고, μ에 대한 95% 신뢰구간은 $\square\ \overline{y} \pm 2.262\,s/\sqrt{n}\ \square$인데, $\overline{y} = \sum_{i=1}^{n} y_i/n$이고 $s^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2/(n-1)$이다.

정규 모분포의 분산 σ^2에 대한 구간 추정시에는 모평균 μ가 알려져 있으면 식 (2.15.10)의 C_n을 PQ로 사용하고, μ를 모르는 경우에는 식 (2.15.11)의 C_{n-1}을 PQ로 사용한다. 후자의 경우가 더 현실적인데, 예를 들어 $n = 10$인 경우에는 카이제곱분포의 확률표로부터 $0.025 = P(C_9 > 19.0228) = P(C_9 < 2.7004)$를 얻으므로

$$0.95 = P(2.7004 < \frac{SS}{\sigma^2} < 19.0228)$$

$$= P(\frac{SS}{19.0228} < \sigma^2 < \frac{SS}{2.7004})$$

가 된다. (비고: $SS = \sum_{i=1}^{10} (Y_i - \overline{Y})^2$.) 따라서, σ^2에 대한 95% 신뢰구간은 □ $\sum_{i=1}^{10} (y_i - \overline{y})^2/19.0228$ 에서 $\sum_{i=1}^{10} (y_i - \overline{y})^2/2.7004$ 까지□ 이다.

정규 모분포가 아닌 (경우 또는 모분포를 모르는) 경우에도 중심극한정리를 이용해서 모평균 μ에 대한 신뢰구간을 얻을 수 있는데, 이에 대표적인 사례인 비율 추정을 §3.1.5에서 다루었다. (비고: 이항분포의 parameter인 p는 *Bernoulli* 모분포의 평균임. §2.1.1 ,§2.1.2 ,§2.8.6 참조.) 또한, 소위 one-sided 신뢰구간이라는 것이 있는데, (지금까지 등장한 two-sided 신뢰구간에 비해서) 잘 쓰이지 않으므로 이를 생략한다.

3.7.2 모집단 두 개의 경우

§2.15.6에서 F분포를 등장시킬 때 두 개의 정규 모분포를 가정했는데, 그때 정의한 확률변수들을 계속 사용한다.

먼저, 식 (2.15.14)의 $(S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$을 PQ로 사용해 보자. 예를 들어, $n_1 = 10$ 이고 $n_2 = 5$ 인 경우에는 F분포의 확률표로부터 $0.025 = P(F_{9,4} > 8.90)$ $= P(F_{9,4} < 1/4.72)$를 얻으므로 (비고: $P(F_{4,9} > 4.72) = 0.025$),

$$0.95 = P(\frac{1}{4.72} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < 8.90)$$

$$= P(\frac{S_2^2}{4.72 S_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{8.90 S_2^2}{S_1^2})$$

가 된다. 따라서, $s_1^2 = \sum_{i=1}^{10} (y_i - \overline{y})^2/9$, $s_2^2 = \sum_{j=1}^{5} (x_j - \overline{x})^2/4$ 라 하면 □ σ_2^2/σ_1^2□에 대한 95% 신뢰구간은 □ $s_2^2/(4.72 s_1^2)$에서 $8.90 s_2^2/s_1^2$까지□이다.

다음, $(\mu_1 - \mu_2)$에 대한 신뢰구간은 다음과 같이 구한다. <비고 2.15.1>에 의해서 $\overline{Y} \sim N(\mu_1, \sigma_1^2/n_1)$이고 $\overline{X} \sim N(\mu_2, \sigma_2^2/n_2)$인데, $\overline{Y}$와 $\overline{X}$가 서로 독립이므로 다시 <비고 2.15.1>에 의해서

$$(\overline{Y}-\overline{X}) \sim N(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})$$

가 된다. 이제, σ_1^2과 σ_2^2이 알려진 경우에는 PQ로

$$Z=\frac{(\overline{Y}-\overline{X})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

를 사용하면 (이후 과정은 앞에서와 동일함), $(\mu_1-\mu_2)$에 대한 95% 신뢰구간으로

$$(\overline{y}-\overline{x})\pm 1.96\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{3.7.2}$$

을 얻는다.

그러나, 더욱 현실적인 경우는 σ_1^2과 σ_2^2을 모르는 경우인데, 조금 복잡하기는 하지만 어차피 4장 이후에 등장할 내용이므로 예습삼아 다룬다. 이제 정규 모분포라는 가정에다가 □$\sigma_1^2=\sigma_2^2$□이라는 가정을 추가한다. 그러면, $\sigma_1^2=\sigma_2^2$이므로 σ_1^2과 σ_2^2을 따로 추정하지 않고 묶어서 한꺼번에 추정하는데, 소위 pooled 추정량이라는 것은 다음과 같다.

$$S^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)} \tag{3.7.3}$$

즉, S^2은 S_1^2과 S_2^2의 가중평균인데, 이때 가중치로는 각각의 자유도를 사용한다.

<비고 3.7.1> 식 (3.7.3)은 식 (2.14.2)의 우변 또는 식 (2.14.8)의 $E[V(Y_1 \mid Y_2)]$에 해당된다. 단, 차이점은 식 (3.7.3)은 표본 통계량이고 $E[V(Y_1 \mid Y_2)]$는 모집단에 관련된 것이라는 점이다.

식 (3.7.1)의 σ_1^2과 σ_2^2을 식 (3.7.3)의 S^2으로 대체하면 PQ로

$$T_{n_1+n_2-2}=\frac{(\overline{Y}-\overline{X})-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{3.7.4}$$

을 얻는데, 이는 자유도가 $(n_1 + n_2 - 2)$인 t분포를 따른다.

<비고 3.7.2> $(n_1 + n_2 - 2)S^2/\sigma^2$은 자유도가 $(n_1 + n_2 - 2)$인 카이제곱분포를 따르며 (§2.12.6 참조), 정규분포를 따르는 $(\overline{Y} - \overline{X})$와 독립임.

예를 들어, $n_1 + n_2 = 11$ 인 경우에는 $(\mu_1 - \mu_2)$에 대한 95% 신뢰구간으로 식 (3.7.2) 대신에

$$(\overline{y} - \overline{x}) \pm 2.262\ s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

을 얻는다. (단, s는 식 (3.7.4)의 S에 대응하는 표본 관찰치임.)

정규 모분포가 아닌 (경우 또는 모분포를 모르는) 경우에도 중심극한정리를 이용해서 $(\mu_1 - \mu_2)$에 대한 근사적 신뢰구간을 얻을 수 있다. 이때, 식 (3.7.1)을 PQ로 사용하는데, 유의할 점은 다음과 같다. σ_1^2과 σ_2^2을 모르는 경우에 이들은 식 (3.7.1)에서는 각각의 추정량으로 그리고 식 (3.7.2)에서는 각각의 추정치로 대체하더라도 식 (3.7.1)은 여전히 근사적으로 $N(0,1^2)$이고 (<비고 2.15.3> 참조), 따라서 식 (3.7.2)는 여전히 유효하다. 예를 들어, 모평균이 각각 p_1과 p_2인 두 개의 *Bernoulli* 모집단에서 크기가 각각 n_1과 n_2인 표본을 (복원) 추출해서 얻은 최우추정치를 각각 $\widehat{p_1}$과 $\widehat{p_2}$이라 하면, $(p_1 - p_2)$에 대한 95% 신뢰구간은 다음과 같다.

$$(\widehat{p_1} - \widehat{p_2}) \pm 1.96\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}} \qquad (3.7.5)$$

즉, 식 (3.7.2)에서 $\overline{y}$와 $\overline{x}$는 각각 $\widehat{p_1}$과 $\widehat{p_2}$에 해당되는데, $\sigma_1^2 = p_1(1-p_1)$과 $\sigma_2^2 = p_2(1-p_2)$는 알려지지 않은 값이므로 이들을 각각 $\widehat{p_1}(1-\widehat{p_1})$과 $\widehat{p_2}(1-\widehat{p_2})$으로 대체한 것이다 (식 (3.1.13) 참조).