4. 분산의 신뢰구간

<정리 6.4-1> $X_1, X_2, \cdots, X_n$을 정규분포 $N(\mu, \sigma^2)$으로부터 추출한 확률표본이라고 할 때, 평균 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$와 분산 $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$사이에는 다음의 성질이 성립한다.

(1) $\overline{X}$와 S^2은 서로 독립이다.

(2) $\frac{nS^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})}{\sigma^2}$는 자유도가 $n-1$인 카이제곱분포 $\chi^2(n-1)$를 따른다.

<증명> (1) By Basu 정리

< Basu 정리> 통계량 T와 θ의 완비충분 통계량이고 $Y = Y(\theta)$와 독립이면 T와 Y믐 θ에 대하여 서로 독립이다.

<충분통계량(6.6절)> $X_1, X_2, \cdots, X_n$: 모수가 θ인 분포에서의 확률표본, $T = g(X_1, X_2, \cdots, X_n)$은 θ의 추정량이라고 하자. $T = t$가 주어졌을 때, 확률 $P[X_1 = x_1, \cdots, X_n = x_n \mid T = t]$가 θ와 무관한 경우에 통계량 T는 θ의 충분통계량이라고 한다.

• 통계량 조건부분포가 모수 θ에 의존하지 않는다는 것은 통계량 T가 이미 θ에 대한 모든 정보를 설명하고 있기 때문에, T가 주어지고 나머지 정보(n개의 표본)는 θ를 설명하는 데 아무런 도움을 주지 못한다는 뜻이다.

<완비성> $X_1, X_2, \cdots, X_n$: 모수가 θ인 분포에서의 확률표본, $T = g(X_1, X_2, \cdots, X_n)$에 대하여, $E(\varphi(T)) = 0$을 만족하는 θ에 무관한 함수 φ가 $\varphi(\cdot) = 0$뿐이라면, T를 완비통계량이라고 한다. 만약 T가 θ에 대한 충분통계량이라면, 이를 완비 충분통계량이라 한다.

(2) $\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X} + \overline{X} - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2 + \frac{n(\overline{X} - \mu)^2}{\sigma^2} \quad \cdots (*)$

$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$: 표준정규분포 $N(0,1)$을 따름 $\Rightarrow$ $Z^2 = \frac{n(\overline{X} - \mu)^2}{\sigma^2}$: $\chi^2(1)$을 따름

식$(*)$에서 $W := \frac{nS^2}{\sigma^2} + Z^2$라 하면 $\overline{X}$와 S^2은 (1)에 의하여 서로 독립$\Rightarrow S^2$와 Z^2 역시 독립.

• W의 적률생성함수: $E[e^{tW}] = E\left[e^{t\left(\frac{nS^2}{\sigma^2} + Z^2\right)}\right] = E\left[e^{t(nS^2/\sigma^2)}\right] \cdot E\left[e^{tZ^2}\right]$.

• W의 분포는 $\chi^2(n)$이므로 적률생성함수: $E[e^{tW}] = (1-2t)^{-\frac{n}{2}}$

Z^2의 분포는 $\chi^2(1)$이므로 적률생성함수: $E[e^{tZ}] = (1-2t)^{-\frac{1}{2}}$

$$E\left[e^{t(nS^2/\sigma^2)}\right] = E[e^{tW}]/E\left[e^{tZ^2}\right] = (1-2t)^{-\frac{n-1}{2}},\ t < \frac{1}{2}.$$

$\therefore$ $\frac{nS^2}{\sigma^2}$의 분포는 자유도가 $n-1$인 카이제곱분포 $\chi^2(n-1)$를 따른다.

• 단일 모집단의 경우 σ^2의 신뢰구간: $\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$이므로

$$P\left[\chi^2_{1-\frac{\alpha}{2}}(n-1) \le \frac{nS^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2}}(n-1)\right] = 1 - \alpha.$$

$\therefore \sigma^2$의 $100(1-\alpha)\%$ 신뢰구간: $\left[\frac{nS^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right]$

<정의 6.4-1> U와 V가 서로 독립이고 각각 자유도가 r_1, r_2인 카이제곱분포를 따를 때,

$$F=\frac{U/r_1}{V/r_2}$$

는 분자의 자유도가 r_1, 분모의 자유도가 r_2인 F-분포((F-distribution 또는 Fisher-Snedecor distribution; 분산분석 등에 널리 사용)를 따른다.

• F-분포의 밀도함수

$$f(x) = \frac{\Gamma\left(\frac{r_1+r_2}{2}\right)\left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}} x^{\frac{r_1}{2}-1}}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)\left(1+\frac{r_1}{r_2}x\right)^{\frac{r_1+r_2}{2}}}, \quad o < x < \infty.$$

<정리 6.4-2> F-분포의 평균과 분산

$$E(F)=\frac{r_2}{r_2-2},\quad Var(F)=\frac{2r_2^2(r_1+r_2-2)}{r_1(r_2-2)^2(r_2-4)}$$

<증명> 자유도 r인 카이제곱분포 $\chi^2(r)$의 밀도함수와 평균, 분산, 적률생성함수

$$f(x)=\frac{1}{\Gamma\left(\frac{r}{2}\right)2^{\frac{r}{2}}}x^{\frac{r}{2}-1}e^{-\frac{x}{2}},\ 0\le x<\infty.$$

$$\mu=r,\ \sigma^2=2r,\ M(t)=(1-2t)^{-\frac{r}{2}}.$$

• $E\left(\frac{U}{r_1}\right)=\frac{1}{r_1}E(U)=\frac{1}{r_1}\cdot r_1=1,\ E(U^2)=(E(U))^2+2r_1=r_1^2+2r_1$

• $E\left(\frac{1}{V}\right)=\int_0^\infty \frac{1}{x}\cdot\frac{1}{\Gamma\left(\frac{r_2}{2}\right)2^{\frac{r_2}{2}}}x^{\frac{r_2}{2}-1}e^{-\frac{x}{2}}dx$

$$=\frac{\Gamma\left(\frac{r_2}{2}-1\right)}{2\Gamma\left(\frac{r_2}{2}\right)}\int_0^\infty\frac{1}{\Gamma\left(\frac{r_2}{2}-1\right)2^{\frac{r_2}{2}-1}}x^{\left(\frac{r_2}{2}-1\right)-1}e^{-\frac{x}{2}}dx=\frac{1}{2\left(\frac{r_2}{2}-1\right)}=\frac{1}{r_2-2}.$$

• $E\left(\frac{1}{V^2}\right)=\int_0^\infty\frac{1}{x^2}\cdot\frac{1}{\Gamma\left(\frac{r_2}{2}\right)2^{\frac{r_2}{2}}}x^{\frac{r_2}{2}-1}e^{-\frac{x}{2}}dx$

$$=\frac{\Gamma\left(\frac{r_2}{2}-2\right)}{2^2\Gamma\left(\frac{r_2}{2}\right)}\int_0^\infty\frac{1}{\Gamma\left(\frac{r_2}{2}-2\right)2^{\frac{r_2}{2}-2}}x^{\left(\frac{r_2}{2}-2\right)-1}e^{-\frac{x}{2}}dx$$

$$=\frac{1}{4\left(\frac{r_2}{2}-1\right)\left(\frac{r_2}{2}-2\right)}=\frac{1}{(r_2-2)(r_2-4)}.$$

$\therefore$ F-통계량의 평균과 2차 적률 및 분산

$$E(F)=E\left(\frac{U/r_1}{V/r_2}\right)=E\left(\frac{U}{r_1}\right)\cdot E\left(\frac{r_2}{V}\right)=\frac{r_2}{r_2-2}.$$

$$E(F)=E\left[\left(\frac{U}{r_1}\right)^2\right]\cdot E\left[\left(\frac{r_2}{V}\right)^2\right]=\frac{(r_1^2+2r_1)}{r_1^2}\cdot\frac{r_2^2}{(r_2-2)(r_2-4)}=\frac{r_2^2(r_1^2+2r_1)}{r_1^2(r_2-2)(r_2-4)}$$

$$Var(F)=E(F^2)-(E(F))^2=\frac{r_2^2(r_1^2+2r_1)}{r_1^2(r_2-2)(r_2-4)}-\frac{r_2^2}{(r_2-2)^2}=\frac{(r_2^2r_1^2+2r_1r_2^2)(r_2-2)-r_2^2(r_1^2(r_2-4))}{r_1^2(r_2-2)^2(r_2-4)}=\frac{2r_2^2(r_1+r_2-2)}{r_1(r_2-2)^2(r_2-4)}$$

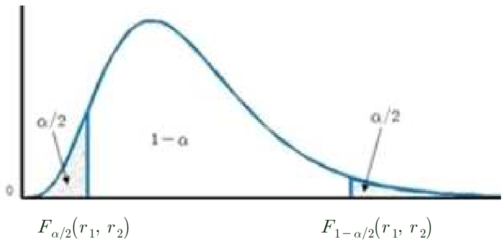
[참고] F-분포의 성질

① $F_{1-\alpha/2}(r_1, r_2) = \dfrac{1}{F_{\alpha/2}(r_2, r_1)}$

② $X_1,\ X_2,\ \cdots,\ X_m \ \sim\ N(\mu_X, \sigma_X^2)$

$Y_1,\ Y_2,\ \cdots,\ Y_n \ \sim\ N(\mu_Y, \sigma_Y^2)$

X_i와 Y_j는 서로 독립이라고 하자.



- $\dfrac{nS_Y^2}{\sigma_Y^2} \sim \chi^2(n-1)$, $\dfrac{nS_X^2}{\sigma_X^2} \sim \chi^2(m-1)$ 서로 독립

$$F = \frac{nS_Y^2/\sigma_Y^2(n-1)}{mS_X^2/\sigma_X^2(m-1)} \sim F(n-1,\ m-1)$$

- 두 모분산의 비율에 대한 신뢰구간

$P[F \le c] = \dfrac{\alpha}{2}$, $P[F \le d] = 1 - \dfrac{\alpha}{2}$를 만족하는 c와 d라 하면

$$P[c \le F \le d] = P\left[c \cdot \frac{m\,S_X^2/(m-1)}{n\,S_Y^2/(n-1)} \le \frac{\sigma_X^2}{\sigma_Y^2} \le d \cdot \frac{m\,\sigma_X^2/(m-1)}{n\,S_Y^2/(n-1)}\right] = 1-\alpha.$$

$\therefore \dfrac{\sigma_X^2}{\sigma_Y^2}$의 $100(1-\alpha)\%$ 신뢰구간

$$\left[c \cdot \frac{m\,S_X^2/(m-1)}{n\,S_Y^2/(n-1)},\quad d \cdot \frac{m\,S_X^2/(m-1)}{n\,S_Y^2/(n-1)}\right].$$

5. 평균의 신뢰구간<모분산을 모르는 경우>

<정의 6.5-1> 확률변수 Z가 표준정규분포 $N(0,1)$을 따르고, 확률변수 U는 자유도가 r인 카이제곱분포 $\chi^2(r)$를 따르며 Z와 U가 서로 독립일 때, 통계량

$$T = \frac{Z}{\sqrt{U/r}}$$

는 자유도가 r인 t-분포 $t(r)$를 따른다. t-분포의 밀도함수

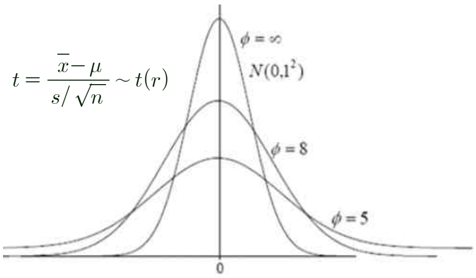
$$f(x) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi r}\,\Gamma\left(\frac{r}{2}\right)\left(1+\frac{x^2}{r}\right)^{\frac{r+1}{2}}}, \quad -\infty < x < \infty.$$

<정리 6.5-1> 자유도가 r인 t-분포 $t(r)$의 통계량 T의 평균과 분산

$$E(T) = 0, \quad Var(T) = E(T^2) = \frac{r}{r-2}, \quad r \geq 2.$$

<증명> $E(T) = E(Z) \cdot E\left(\sqrt{r/U}\right) = 0 \cdot E\left(\sqrt{r/U}\right) = 0$

$$Var(T) = E(T^2) - (E(T))^2 = E(Z^2) \cdot E\left(\frac{r}{U}\right) = 1 \cdot r \cdot E\left(\frac{1}{U}\right) = \frac{r}{r-2}.$$



t-분포는 정규분포와 마찬가지로 좌우대칭인 분포이며 정규분포보다 꼬리 부분의 확률 값이 더 크다. 그러나 자유도가 커지면 t-분포는 정규분포와 거의 같게 된다.

자유도(r)가 ∞이면 t분포는 정규분포와 일치

• 단일 모집단에서 평균의 신뢰구간

$X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2) \Rightarrow \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, $\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$이며 이 둘은 서로 독립

$$T = \frac{\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{\frac{nS^2}{\sigma^2(n-1)}}} = \frac{\overline{X}-\mu}{S/\sqrt{n-1}} \sim t(n-1)$$

$$\begin{aligned} 1-\alpha &= P\left[-t_{\alpha/2}(n-1) \leq \frac{\overline{X}-\mu}{S/\sqrt{n-1}} \leq t_{\alpha/2}(n-1)\right] \\ &= P\left[\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n-1}} \leq \mu \leq \overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n-1}}\right]. \end{aligned}$$

[참고] ① 신뢰구간 $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$의 폭은 $\overline{X} \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{n-1}}$보다 좁다.

② σ^2을 모르는 경우라도 표본의 크기가 약 30 이상으로 큰 경우에는 통계량 T의 분포는 근사적으로 표준정규분포에 가깝게 된다.

• 두 모집단에서 평균의 차에 대한 신뢰구간

$X_1, X_2, \cdots, X_n \sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, \cdots, Y_n \sim N(\mu_Y, \sigma_Y^2)$이며 두 표본은 서로 독립이고 두 표본의 분산은 같다 $(\sigma_X^2 = \sigma_Y^2 = \sigma^2)$고 가정

통계량 $U = \dfrac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma^2}{m} + \dfrac{\sigma^2}{n}}} \sim N(0,1)$

통계량 $V = \dfrac{mS_X^2}{\sigma^2} + \dfrac{mS_Y^2}{\sigma^2} \sim \chi^2(m+n-2)$

U와 V는 서로 독립이므로 통계량

$$T = \frac{\dfrac{\left[\overline{X} - \overline{Y} - (\mu_X - \mu_Y)\right]}{\sqrt{\dfrac{\sigma^2}{m} + \dfrac{\sigma^2}{n}}}}{\sqrt{\dfrac{(mS_X^2 + nS_Y^2)}{\sigma^2(m+n-2)}}}$$

$$= \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\left[(mS_X^2 + nS_Y^2)/(m+n-2)\right]\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}} = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}}$$

$\Rightarrow$자유도가 $m+n-2$인 t-분포 $t(m+n-2)$

여기서 $S_p^2 = \dfrac{mS_X^2 + nS_Y^2}{m+n-2}$; 합동표본분산(Plooled sample variance)

• 두 표본의 평균의 차 $\mu_X - \mu_Y$의 $100(1-\alpha)\%$ 신뢰구간

$$P\left[\overline{X} - \overline{Y} - t_{\alpha/2}(m+n-2)\, S_p\sqrt{\frac{1}{m} + \frac{1}{n}} \le \mu_X - \mu_Y \right.$$
$$\left. \le \overline{X} - \overline{Y} + t_{\alpha/2}(m+n-2)\, S_p\sqrt{\frac{1}{m} + \frac{1}{n}}\right]$$
$$= 1 - \alpha.$$

[참고] 표본의 크기가 클 때, $\mu_X - \mu_Y$의 $100(1-\alpha)\%$ 신뢰구간

$$\overline{X} - \overline{Y} \pm z_{\alpha/2}\sqrt{\frac{S_X^2}{m-1} + \frac{S_Y^2}{n-1}}$$

<예 6.5-1> 서로 독립인 두 집단에서 자료를 얻어 정리한 결과가 다음과 같다고 하자. 평균의 차에 대한 95% 신뢰구간을 구하라.

집단 1 : $m = 15$, $\overline{X} = 781.31$, $s_X^2 = 60.76$.
집단 2 : $n = 9$, $\overline{Y} = 78.61$, $s_Y^2 = 48.24$.

<풀이> $\mu_X - \mu_Y$의 95% 신뢰구간은 자유도가 15+9-2=22인 t-분포표에서 $P[T \le 2.074] = 0.975$이므로

$$81.31 - 78.61 \pm 2.074\sqrt{\frac{9(60.76) + 15(48.24)}{22}\left(\frac{1}{15} + \frac{1}{9}\right)}$$

이다. 즉, $[-3.95,\ 9.35]$이다.

6. 최소분산 불편추정량과 충분통계량

• 좋은 추정량이란 일반적으로 충분성을 갖추고 있고 분산이 작은 통계량

<정의 6.5-1> 최소분산 불편추정량

$X_1, X_2, \cdots, X_n$은 밀도함수가 $f(x:\theta)$인 분포에서의 확률표본이라고 하자. 다음의 두 가지 조건을 만족하는 통계량 T를 θ의 최소분산 불편추정량(Minimum variance unbiased estimator : MVUE)이라고 한다.

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

(b) 다른 어떠한 불편 추정량 T'에 대해서도 $Var(T)\le Var(T')$.

<예 6.6-1> X_1, X_2, X_3를 정규분포 $N(\mu,\sigma^2)$, $-\infty<\mu<\infty$, $\sigma^2>0$에서의 확률표본이라고 하면 표본평균 $\overline{X}$의 기댓값은 $E(\overline{X})=\mu$, 분산은 $Var(\overline{X})=\dfrac{\sigma^2}{n}$이다. 이제 $Y=\dfrac{X_1+2X_2+3X_3}{6}$이라 두면

Y의 기댓값: $E(Y)=\dfrac{1}{6}E(X_1+2X_2+3X_3)=\dfrac{1}{6}(\mu+2\mu+3\mu)=\mu$.

Y의 분산: $Var(Y)=\left(\dfrac{1}{6}\right)^2\sigma^2+\left(\dfrac{2}{6}\right)^2\sigma^2+\left(\dfrac{3}{6}\right)^2\sigma^2=\dfrac{7}{18}\sigma^2$

$\therefore$ $\overline{X}$와 Y는 μ의 불편 추정량이며 $Var(\overline{X})<Var(Y)$임을 알 수 있다.

<예 6.6-2> $X_1, X_2, \cdots, X_n$을 정규분포 $N(\mu,\sigma^2)$에서의 확률표본이라고 하자.

표본분산 $S^2=\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$으로 두면, $E\left[\dfrac{n}{n-1}S^2\right]=\sigma^2$이며 $E[(cS^2-\sigma^2)^2]$을 최소로 하는 상수 c (cS^2의 형태를 갖는 σ^2의 최소평균 제곱오차 추정량(Minimum mean square error estimator))를 구해보자

<풀이> $W=cS^2-\sigma^2$, $E[W^2]=\mu_W^2+\sigma_W^2$이고,

$$E(S^2)=\frac{n-1}{n}\sigma^2,\quad Var(S^2)=\frac{2(n-1)}{n^2}\sigma^4\text{이므로}$$

$$E[(cS^2-\sigma^2)^2]=[E(cS^2-\sigma^2)]^2+Var(cS^2-\sigma^2)$$

$$=\left[c\frac{n-1}{n}\sigma^2-\sigma^2\right]^2+\frac{2c^2(n-1)}{n^2}\sigma^4.$$

$$\frac{\partial E[(cS^2-\sigma^2)^2]}{\partial c}=0\ \Rightarrow 2\left[c\frac{n-1}{n}\sigma^2-\sigma^2\right]\frac{n-1}{n}\sigma^2+4c\frac{n-1}{n^2}\sigma^4=0$$

$$\therefore\ c=\frac{n}{n+1}.$$

σ^2의 최소제곱오차 추정량

$\dfrac{n}{n+1}S^2=\dfrac{1}{n+1}\sum_{i=1}^{n}(X_i-\overline{X})^2$. [참고] S^2은 σ^2의 최우 추정량이고, $\dfrac{n}{n-1}S^2$은 σ^2의 불편 추정량이다.

<예 6.6-3> $X_1, X_2, \cdots, X_n$을 정규분포 $N(\mu,\sigma^2)$에서의 확률표본이라고 하자.

$cS=c\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}$이 σ의 불편추정량이 되도록 상수 c의 값을 구하여라.

<풀이> $\dfrac{nS^2}{\sigma^2}\sim\chi^2(n-1)$, $Y=\dfrac{nS^2}{\sigma^2}$라 두면

$$E[\sqrt{Y}]=\int_0^\infty\sqrt{y}\cdot\frac{1}{\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}}y^{\frac{n-1}{2}-1}e^{-\frac{y}{2}}dy$$

$$=\frac{\Gamma\left(\frac{n}{2}\right)\sqrt{2}}{\Gamma\left(\frac{n-1}{2}\right)}\int_0^\infty\frac{1}{\Gamma\left(\frac{n}{2}\right)2^{\frac{n}{2}}}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}dy=\frac{\Gamma\left(\frac{n}{2}\right)\sqrt{2}}{\Gamma\left(\frac{n-1}{2}\right)}$$

$$E[S]=E\left[\sqrt{\frac{\sigma^2}{n}\cdot\frac{nS^2}{\sigma^2}}\right]=\frac{\sigma}{\sqrt{n}}E(\sqrt{Y})=\frac{\sqrt{2}}{\sqrt{n}}\cdot\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\cdot\sigma$$

$$\therefore c=\frac{\sqrt{n}}{\sqrt{2}}\cdot\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

※ 라오-블랙웰의 정리(Rao-Blackwell theorem)를 이용하면 보다 간편하게 최소분산 불편추정량을 찾을 수 있다.

<정리 6.6-1> 라오-블랙웰의 정리(Rao-Blackwell theorem)
U가 T의 통계량이고 $E(U)=\theta$ (즉, U는 θ의 불편 추정량)라 하자. $W=E(U|T)$라 두면 다음 식이 성립한다.
(1) W는 T의 함수이다.
(2) $E(W)=\theta$.
(3) $Var(W)\le Var(U)$.
<증명> (1) W는 T가 주어졌을 때 U의 기댓값이므로 T의 함수가 된다.
(2) $E(W)=E[E(U|T)]=E(U)=\theta$, by95쪽 이중 기댓값 정리 $E(X)=E[E(X|Y)]$
(3) $Var(U)=E[U-\theta]^2=E[U-W+W-\theta]^2$

$$=E[U-W]^2+2E[(U-W)(W-\theta)]+E[W-\theta]^2$$

$$E[(U-W)(W-\theta)]=E[E[(U-W)(W-\theta)|T]]=(W-\theta)E[(U-W)|T]$$

$$\xrightarrow{E(U|T)=W}(W-\theta)[W-E[W|T]]=0$$

$$Var(U)=E[U-W]^2+E[W-\theta]^2=E[U-W]^2+Var(W)\ge Var(W)$$

<정의 6.6-2> 충분 통계량 (Sufficient statistic)
$X_1, X_2, \cdots, X_n$을 모수 θ인 분포에서의 확률표본이라 하고 $T=g(X_1, X_2, \cdots, X_n)$은 θ의 추정량이라고 하자. T의 값 $T=t$가 주어졌을 때, 확률 $P[X_1=x_1, \cdots, X_n=x_n|T=t]$가 θ와 무관한 경우에 통계량 T는 θ의 충분 통계량이라고 한다.
<예 6.6-4> $X_1, X_2, \cdots, X_n$을 베르누이 시행 $B(1,p)$에서의 확률표본이라고 하자. $P(X_i=\text{성공})=p$이고, $P(X_i=\text{실패})=1-p$이다. 이제 $Y=\sum X_i$라 두면

$$P[X_1=x_1, \cdots, X_n=x_n \mid Y=y] = \frac{P[X_1=x_1, \cdots, X_n=x_n, Y=y]}{P[Y=y]}$$

$$= \frac{p^y(1-p)^{n-y}}{\binom{n}{y}p^y(1-p)^{n-y}} = \frac{1}{\binom{n}{y}}$$

이므로 p와는 무관하다. 따라서 $Y=\sum X_i$는 p의 충분 통계량이다.

[참고] ① $P[X_1=x_1, \cdots, X_n=x_n \mid X_1=x_1, \cdots, X_n=x_n]=1$이므로 모수 θ 와는 무관하다. 즉, 확률표본 자체는 θ의 충분 통계량이다.
② $Y_1, \cdots, Y_n$을 $X_1, \cdots, X_n$의 순서 통계량이라고 하자.

$$P[X_1=x_1, \cdots, X_n=x_n \mid Y_1=y_1, \cdots, Y_n=y_n]=\frac{1}{n!}$$

이므로 모수 θ와는 무관하다. 따라서 순서 통계량은 θ의 충분 통계량이다.

<예 6.6-5> X_1, X_2는 밀도함수가 $f(x:\lambda)=\dfrac{\lambda^x e^{-\lambda}}{x!}$, $x=0, 1, \cdots$인 포아송 분포에서의 확률표본이라고 하자. X_1-X_2는 λ의 충분 통계량인가?
<풀이> $P[X_1=0, X_2=0 \mid X_1-X_2=0] = P[X_1=0, X_2=0 \mid X_1=X_2]$

$$=\frac{P[X_1=0, X_2=0, X_1=X_2]}{P[X_1=X_2]} = \frac{P[X_1=0, X_2=0]}{P[X_1=X_2]}$$

$\langle P[X_1=X_2] = \sum_{k=0}^{\infty} P[X_1=k, X_2=k] = \sum_{k=0}^{\infty} P[X_1=k]P[X_2=k]$

$$= \sum_{k=0}^{\infty}\left(e^{-\lambda}\frac{\lambda^x}{x!}\right)^2 = e^{-2\lambda}\left[1+\lambda^2+\left(\frac{\lambda^2}{2!}\right)^2+\cdots\right]\text{이므로}\rangle$$

$$= \frac{(f(0:\lambda))^2}{\sum_{k=0}^{\infty}(f(k:\lambda))^2} = \frac{e^{-2\lambda}}{e^{-2\lambda}\left[1+\lambda^2+\left(\frac{\lambda^2}{2!}\right)^2+\cdots\right]} = \frac{1}{1+\lambda^2+\left(\frac{\lambda^2}{2!}\right)^2+\left(\frac{\lambda^3}{3!}\right)^2+\cdots}$$

이므로 λ가 포함되어 있다. 따라서, X_1-X_2는 λ의 충분 통계량이 될 수 없다.

<정리 6.6-2> 피셔-네이만의 인수분해 정리 (Fisher-Neyman's factorization theorem))
U가 확률표본 $X_1, X_2, \cdots, X_n$으로 만들어진 통계량이라고 할 때, U가 모수 θ의 충분 통계량이 될 필요충분조건은 우도함수가 다음과 같이 인수분해되는 것이다.

$$L(\theta; x_1, \cdots, x_n) = g(u:\theta) \cdot h(x_1, \cdots, x_n).$$

<예 6.6-6> $X_1, \cdots, X_n$는 밀도함수가 $f(x:p) = p^x(1-p)$, $x = 0, 1, \cdots$인 기하분포에서의 확률표본이라고 하면 우도함수는

$$L(p: x_1, \cdots, x_n) = p^{\sum x_i}(1-p)^n$$

이 된다. $u = \sum x_i$이라 하고 $g(u, p) = p^u(1-p)^n$, $h(x_1, \cdots, x_n) = 1$로 두면 위의 우도함수는 $g(u, \theta)$와 $h(x_1, \cdots, x_n)$의 곱으로 인수분해된다. 따라서 $U = \sum X_i$는 θ의 충분 통계량이 된다.

<예 6.6-7> $X_1, \cdots, X_n$을 정규분포 $N(0, \sigma^2)$에서의 확률표본이라고 하고, $\theta = \sigma^2$으로 두면 우도함수는

$$L(\theta: x_1, \cdots, x_n) = \left(\frac{1}{\sqrt{2\pi\theta}}\right)^n \exp\left(-\frac{1}{2\theta}\sum x_i^2\right)$$

이다. $u = \sum x_i^2$이라 하고 $g(u, \theta) = \left(\frac{1}{\sqrt{2\pi\theta}}\right)^n \exp\left(-\frac{1}{2\theta}\sum x_i^2\right)$, $h(x_1, \cdots, x_n) = 1$로 두면 위의 우도함수는 $g(u, \theta)$와 $h(x_1, \cdots, x_n)$의 곱으로 인수분해된다. 따라서 $U = \sum X_i^2$는 θ의 충분 통계량이 된다.

<예 6.6-8> $X_1, \cdots, X_n$을 균일분포 $u(0, \theta)$에서의 확률표본이라고 하면 우도함수

$$L(\theta: x_1, \cdots, x_n) = \begin{cases} \frac{1}{\theta^n}, & 0 \le x_{(n)} \le \theta \\ 0, & \text{그 외} \end{cases}$$

이 된다. $u = x_{(n)}$이라고 하면 위의 우도함수는

$$g(u: \theta) = \begin{cases} \frac{1}{\theta^n}, & 0 \le x_{(n)} \le \theta \\ 0, & \text{그 외} \end{cases}$$

와 $h(x_1, \cdots, x_n) = 1$의 곱으로 인수분해된다. 따라서 $U = X_{(n)}$는 θ의 충분 통계량

[참고] 위의 인수분해정리와 최우(최대가능도) 추정량의 정의로부터 최우 추정량은 충분 통계량의 함수로 주어짐을 알 수 있다.

<따름정리> 정리 6.6-1의 따름정리

$E(U)=\theta$이고, 통계량 T가 θ의 충분 통계량이면 $W=E[U|T]$는 충분 통계량의 함수로 주어지며 θ의 불편 추정량이 되고 W의 분산은 U의 분산보다 크지 않다. 즉, $Var(W)\le Var(U)$이다.

<예 6.6-9> $X_1, \cdots, X_n$을 서로 독립인 베르누이 시행 $B(1,p)$에서의 확률표본이라고 하면 $\sum X_i$는 p의 충분 통계량이고 $E(X_1)=p$이므로 $U=X_1$, $T=\sum X_i$라 두면 위의 따름정리로부터 $E[U|T]$는 T의 함수이며 p의 불편추정량이다.

$$E\left[X_1 \mid \sum_{i=0}^{1} X_i = y\right] = \sum_{x_1=0}^{1} x_1 \cdot P\left[X_1 = x_1 \mid \sum_{i=1}^{n} X_i = y\right] = 1 \cdot P\left[X_1 = 1 \mid \sum_{i=1}^{n} X_i = y\right]$$

$$= \frac{P\left[X_1 = 1,\ \sum_{i=1}^{n} X_i = y\right]}{P\left[\sum_{i=1}^{n} X_i = y\right]} = \frac{P\left[X_1 = 1,\ \sum_{i=2}^{n} X_i = y-1\right]}{P\left[\sum_{i=1}^{n} X_i = y\right]}$$

$$= \frac{p \cdot \binom{n-1}{y-1} p^{y-1}(1-p)^{n-1-y+1}}{\binom{n}{y} p^{y}(1-p)^{n-y}}$$

$$= \frac{\binom{n-1}{y-1}}{\binom{n}{y}} = \frac{y}{n} = \sum \frac{X_i}{n} = \overline{X}.$$

또한 $\sum X_i$의 분포는 $B(n,p)$이므로 $E(\overline{X})=p$, $Var(\overline{X})=\dfrac{p(1-p)}{n}$이다. 앞의 따름정리에 따라서 $U=\overline{X}$, $T=\sum X_i$라 두면

$$E[\overline{X} \mid \sum X_i = y] = E\left[\frac{1}{n}\sum X_i \mid \sum X_i = y\right] = \frac{y}{n}$$

임을 알 수 있다. 따라서 $Var(\overline{X})=\dfrac{p(1-p)}{n} \le p(1-p)= Var(X_1)$이므로 $\overline{X}$는 p의 최소분산 불편추정량임을 알 수 있다.

<예 6.6-10> $X_1, \cdots, X_n$을 균일분포 $U(0,\theta)$에서의 확률표본이라고 하면 $X_{(n)}$은 θ의 충분 통계량이고, $E[X_{(n)}]=\dfrac{n}{n+1}\theta$이므로 $\dfrac{n+1}{n}X_{(n)}$는 θ의 불편 추정량이다.

$$E\left[\frac{n+1}{n} X_{(n)} \mid X_{(n)} = y\right] = \frac{n+1}{n} y$$

이므로 $\dfrac{n+1}{n}X_{(n)}$는 θ의 최소분산 불편추정량이 된다.

<정리 6.6-3> 크레이머-라오의 하한 (cramer-Rao lower bound; CRLB)
$X_1, \cdots, X_n$은 밀도함수가 $f(x:\theta)$인 분포에서의 확률표본이라 하고 모수 θ는 $\theta \in \Omega = \{\theta : c < \theta < d\}$인 관계가 있으며 c와 d는 θ에 무관하다고 하자. $\frac{\partial}{\partial\theta}\ln f(x:\theta)$의 값이 존재하면 θ의 불편 추정량 U에 대하여 다음의 부등식이 성립한다.

$$Var(U) \geq \frac{1}{I(\theta)}.$$

여기서 $I(\theta)$는 피셔의 정보량 (Fisher's information)

$$I(\theta) = nE\left[\left(\frac{\partial}{\partial\theta}\ln f(x:\theta)\right)^2\right] = -nE\left[\frac{\partial^2}{\partial\theta^2}\ln f(x:\theta)\right]$$

• 모수 θ의 불편 추정량 T의 분산이 $\frac{1}{I(\theta)}$(=CRLB)와 같다면 추정량 T는 최소분산 불편추정량(MVUE)이 된다. 피셔의 정보량 $I(\theta)$는 표본의 크기가 클수록 증가하고 분산이 클수록 감소한다.

<예 6.6-11> $X_1, \cdots, X_n$을 밀도함수가 $f(x:p) = p^x(1-p)^{1-x}$, $x = 0, 1$인 베르누이 시행 $B(1,p)$에서의 확률표본이라고 하자. 표본 평균 $\overline{X}$의 CRLB를 구해보자

$$\frac{\partial}{\partial p}\ln f(x:p) = \frac{\partial}{\partial p}(x\ln p + (1-x)\ln(1-p)) = \frac{x}{p} - \frac{1-x}{1-p}, \quad \frac{\partial^2}{\partial p^2}\ln f(x:p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

$$= -\frac{1}{(1-p)^2} - \left(\frac{1}{p^2} - \frac{1}{(1-p)^2}\right)x$$ 이고 $E(X) = p$이므로

$$E\left[\frac{\partial^2}{\partial p^2}\ln f(x:p)\right] = -\frac{1}{(1-p)^2} - \left(\frac{1}{p^2} - \frac{1}{(1-p)^2}\right)p = -\frac{1}{p(1-p)}$$ 이다.

따라서 $I(p) = \frac{n}{p(1-p)}$이고 $CRLB = \frac{p(1-p)}{n}$이다. 한편

$Var(\overline{X}) = \frac{p(1-p)}{n} = \frac{1}{I(p)}$이므로 $\overline{X}$는 p의 MVUE가 된다.

<예 6.6-12> $X_1, X_2, \cdots, X_n$을 정규분포 $N(\theta, 1)$에서의 확률표본이라고 할 때, θ의 MVUE를 구하도록 한다.

$$\frac{\partial}{\partial\theta}\ln f(x:\theta) = \frac{\partial}{\partial\theta}\ln\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x-\theta)^2\right)\right] = \frac{\partial}{\partial\theta}\left(-\frac{1}{2}(x-\theta)^2\right) = x-\theta,$$

$$\frac{\partial^2}{\partial\theta^2}\ln f(x:\theta) = \frac{\partial}{\partial\theta}(x-\theta) = -1$$

이므로 피셔의 정보량 $I(\theta)$는 다음과 같이 얻어진다.

$$I(\theta) = -nE\left(\frac{\partial^2}{\partial\theta^2}\ln f(x:\theta)\right) = -nE(-1) = n.$$

따라서 θ의 CRLB는 $\frac{1}{n}$이며, $Var(\overline{X}) = \frac{1}{n} = CRLB$이므로 $\overline{X}$는 θ의 MVUE가 된다.

<예 6.6-13> $X_1, X_2, \cdots, X_n$을 정규분포 $N(0,\sigma^2)$에서의 확률표본이라고 할 때, $\theta=\sigma^2$의 MVUE를 구해보자

$$\begin{aligned}\frac{\partial}{\partial\theta}\ln f(x:\theta) &= \frac{\partial}{\partial\theta}\ln\left[\frac{1}{\sqrt{2\pi\theta}}\exp\left(-\frac{1}{2}\frac{x^2}{\theta}\right)\right] \\ &= \frac{\partial}{\partial\theta}\left[-\frac{1}{2}\ln(2\pi\theta)-\frac{1}{2}\frac{x^2}{\theta}\right] \\ &= -\frac{1}{2\theta}+\frac{x^2}{2\theta^2},\end{aligned}$$

$$\frac{\partial^2}{\partial\theta^2}\ln f(x:\theta) = \frac{\partial^2}{\partial\theta^2}\left(-\frac{1}{2\theta}+\frac{x^2}{2\theta^2}\right)' = \frac{1}{2\theta^2}-\frac{x^2}{\theta^3},$$

$$\begin{aligned}\therefore\ E\left[\frac{\partial^2}{\partial\theta^2}\ln f(x:\theta)\right] &= \frac{1}{2\theta^2}-\frac{1}{\theta^3}E(X^2) \\ &= \frac{1}{2\theta^2}-\frac{1}{\theta^3}\cdot\theta = -\frac{1}{2\theta^2},\end{aligned}$$

$$I(\theta) = -nE\left[\frac{\partial^2}{\partial\theta^2}\ln f(x:\theta)\right] = \frac{n}{2\theta^2}.$$

따라서 θ의 CRLB는 $\frac{2\theta^2}{n}=\frac{2\sigma^4}{n}$이다. 한편 $\widehat{\sigma^2}=\frac{1}{n}\sum X_i^2$로 두면 $E(X_i^2)=\sigma^2$이므로

$$E(\widehat{\sigma^2}) = E\left[\frac{1}{n}\sum X_i^2\right] = \frac{1}{n}\sum E(X_i^2) = \sigma^2$$

즉, $\widehat{\sigma^2}$는 σ^2의 불편 추정량이다. 또한 X_i의 분포가 $N(0,\sigma^2)$이므로 $\frac{X_i^2}{\sigma^2}$의 분포는 $\chi^2(1)$이다. 따라서 $Var\left(\frac{X_i^2}{\sigma^2}\right)=2$, $Var(X_i^2)=2\sigma^4$이다.

$\therefore\ Var(\widehat{\sigma^2})=Var\left(\frac{1}{n}\sum X_i^2\right)=\frac{1}{n^2}\sum Var(X_i^2)=\frac{2\sigma^4}{n}=CRLB.$

그러므로 $\widehat{\sigma^2}=\frac{1}{n}\sum X_i^2$는 σ^2의 MVUE가 된다.

7. 추정량의 기타 성질

X: 동전을 n회 독립적으로 던지는 실험에서 앞면이 나오는 회수

p: 앞면이 나오는 비율, p의 추정량 $\frac{X}{n}$의 분산 $\frac{p(1-p)}{n}$ $(:= h(p))$의 추정량

$Z = h\left(\frac{X}{n}\right) = \frac{X}{n}\left(1-\frac{X}{n}\right)/n$으로 두면

$$\begin{aligned} E(Z) &= E\left\{\frac{X}{n}\left(1-\frac{X}{n}\right)/n\right\} = \frac{1}{n^3}E(nX - X^2) \\ &= \frac{1}{n^3}\{nE(X) - E(X^2)\} \\ &= \frac{1}{n^3}\{n\cdot np - \{(np)^2 + np(1-p)\}\} \\ &= \frac{1}{n^2}(n-1)p(1-p) = \frac{n-1}{n}\frac{p(1-p)}{n} \end{aligned}$$

이므로 Z는 $h(p) = \frac{p(1-p)}{n}$의 편의 추정량이다. 표현을 바꾸면 $\frac{n}{n-1}Z = \frac{n}{n-1}h\left(\frac{X}{n}\right)$은 $h(p)$의 불편 추정량이 된다.

<정리 6.7-1> $h(\theta)$가 θ의 연속 함수이고 $\hat{\theta}$가 θ의 일치 추정량일 때, $h(\hat{\theta})$는 $h(\theta)$의 일치 추정량이 된다.

<증명> 연속함수의 정의 (임의의 $\epsilon > 0$에 대하여 $|\hat{\theta}-\theta| < \delta$에 따라서 $|h(\hat{\theta}) - h(\theta)| < \epsilon$이 되는 $\delta > 0$이 존재한다.) 에 의하여

$$P[|\hat{\theta} - \theta| < \delta] \le P[|h(\hat{\theta}) - h(\theta)| < \epsilon]$$

이 성립하는데 $\hat{\theta}$이 θ의 일치 추정량이므로 $P[|\hat{\theta}-\theta| < \delta] = 1$이다. 따라서 $P[|h(\hat{\theta}) - h(\theta)| < \epsilon] = 1$이 성립한다. 즉 $h(\hat{\theta})$는 $h(\theta)$의 일치 추정량이 된다.

<예 6.7-1> 확률변수 X의 분포가 정규분포 $N(\mu, \sigma^2)$일 때, $P[c < X]$의 일치 추정량을 구해보자.

$$P[c < X] = P\left[\frac{c-\mu}{\sigma} < \frac{X-\mu}{\sigma}\right] = 1 - \Phi\left(\frac{c-\mu}{\sigma}\right)$$

이며 μ와 σ^2이 미지의 모수이면 μ의 최우 추정량 $\hat{\mu} = \overline{X}$와 σ^2의 최우추정량 $\widehat{\sigma^2} = S^2$을 사용하여

$$Z = 1 - \Phi\left(\frac{c-\overline{X}}{S}\right)$$

가 $P[c < X]$의 일치 추정량이 된다.

[참고] $T = \frac{\overline{X}-c}{S/\sqrt{n-1}} = \frac{\sqrt{n}(\overline{X}-c)/\sigma}{\sqrt{\sum(X_i-\overline{X})^2/(n-1)\sigma^2}}$ 는 자유도가 $n-1$인 t-분포 $t(n-1)$을 하는 데 $c \ne \mu$인 경우에 T는 비중심 t-분포 (Noncentral t-distribution)를 한다고 한다.

<예 6.7-2> 확률변수 X의 분포가 정규분포 $N(\mu_X, \sigma^2)$이고 확률변수 Y의 분포가 $N(\mu_Y, \sigma^2)$이며 X와 Y가 서로 독립이고 σ^2을 모르고 있을 때, $P[X < Y]$의 일치 추정량을 구해보자.

<풀이> $X - Y$의 분포는 $N(\mu_X - \mu_Y, 2\sigma^2)$이므로

$$P[X < Y] = P[X - Y < 0] = P\left[\frac{X-Y-(\mu_X-\mu_Y)}{\sqrt{2}\sigma} < \frac{\mu_Y - \mu_X}{\sqrt{2}\sigma}\right]$$

이다. 이제 $X_1, \cdots, X_m$을 정규분포 $N(\mu_X, \sigma^2)$에서의 확률표본이라 하고 $Y_1, \cdots, Y_n$을 정규분포 $N(\mu_Y, \sigma^2)$에서의 확률표본이라고 하면 우도함수는

$$L(\mu_X, \mu_Y, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{m+n}{2}} \exp\left[-\frac{\sum_{i=1}^{m}(X_i-\mu_X)^2 + \sum_{i=1}^{n}(Y_i-\mu_Y)^2}{2\sigma^2}\right]$$

이 된다. 이제 최우 추정량 $\widehat{\mu_X}=\overline{X}$, $\widehat{\mu_Y}=\overline{Y}$, $\widehat{\sigma^2}=\dfrac{mS_X^2+nS_Y^2}{m+n}$을 사용하면 $P[X<Y]$의 일치 추정량은 $\Phi(cT)$의 형태로 표현되며 c와 T는 아래와 같다.

$$c=\sqrt{\frac{(m+n)^2}{2mn(m+n-2)}}, \quad T=\frac{\overline{Y}-\overline{X}}{\sqrt{\dfrac{mS_X^2+nS_Y^2}{m+n-2}\left(\dfrac{1}{m}+\dfrac{1}{n}\right)}}.$$

※ T는 자유도가 $m+n-2$인 t-분포를 따른다.