Daeyoung Lim

Grad School

Year of 2013, late

1. (25점) Let X and Y be independent non-negative random variables with continuous density functions $f_X(x)$ and $f_Y(y)$ respectively on $(0,\infty)$. Show that

   (a) If, given $X+Y=u$, X is uniformly distributed on $(0,u)$ whatever the value of u, then

   $$f_Y(u-v)\,f_X(v)=\frac{1}{u}\int_0^u f_Y(u-y)\,f_X(y)\,dy.$$

   (b) If X and Y be independent exponential random variables with a common parameter λ, $X,Y\sim\mathrm{Exp}(\lambda)$, then the conditional distribution of X given $X+Y=u$ is a uniform distribution on $(0,u)$.

Solution:

(a) 처음 보고 풀기 조금 어려운 문제인 듯하다. 좌변을 먼저 보면, $X+Y=U$와 $X=V$의 결합분포를 구한 것이다. 즉,

$$\begin{cases}U=X+Y\\V=X\end{cases}\implies\begin{cases}X=V\\Y=U-V\end{cases}\tag{1}$$

$$|J|=1\tag{2}$$

$$f_{U,V}(u,v)=f_{X,Y}(v,u-v)\cdot 1\tag{3}$$

$$=f_Y(u-v)\,f_X(v)\tag{4}$$

그리고 우변은 U,V의 분포를

$$f_{V|U}(v\,|\,u)\,f_U(u)\tag{5}$$

의 꼴로 구한 것이다. 문제에서 $V\,|\,U\sim\mathrm{Unif}(0,u)$라고 했고

$$f_U(u)=\int_0^u f_{U,V}(u,v)\,dv\tag{6}$$

$$=\int_0^u f_Y(u-v)\,f_X(v)\,dv\tag{7}$$

이므로

$$f_{V|U}(v|u)\,f_U(u)=\frac{1}{u}\int_0^u f_Y(u-v)\,f_X(v)\,dv\tag{8}$$

이다. 둘 모두 U,V의 결합분포이므로 문제에서 주어진 등식이 성립한다.

(b) (a)에서 한 것을 바탕으로 계산하면

$$f_{U,V}(u,v) = f_Y(u-v)\, f_X(v) \tag{9}$$
$$= \lambda e^{-\lambda(u-v)} \cdot \lambda e^{-\lambda v} \tag{10}$$
$$= \lambda^2 e^{-\lambda u} \tag{11}$$

따라서 $V \mid U$는 U, V의 결합분포에서 U의 주변분포를 나눠야 하므로

$$f_{V|U}(v|u) = f_{U,V}(u,v) \bigg/ \int_0^u f_{U,V}(u,v)\, dv \tag{12}$$
$$= \frac{1}{u} \tag{13}$$

그러므로 $V \mid U \sim \mathrm{Unif}(0,u)$이다.

2. (25점) $X_1, X_2, \ldots, X_n$을 $\mathrm{Unif}(0,\theta)$로부터 얻은 랜덤표본이라고 하자.

(a) 모수 θ의 최대가능도 추정량을 구하라.

(b) 위의 (a)에서 구한 모수 θ의 최대가능도 추정량이 완비충분통계량임을 보여라.

(c) 모수 θ에 대한 $(1-\alpha_1-\alpha_2) \times 100\%$ 신뢰구간을 구하라.

Solution:

(a) 이런 거 물어보지 마.

$$\widehat{\theta} = \max_{1 \le i \le n} X_i \quad \left(= X_{(n)}\right) \tag{14}$$

(b) 완비통계량의 정의상 어떤 통계량 T가 있을 때, 임의의 $\theta \in \Omega$에 대해서

$$\mathrm{E}(r(T)) = 0 \implies \Pr(r(T) = 0) = 1 \tag{15}$$

이어야 하므로 우리의 통계량 $X_{(n)}$의 분포로부터 어떤 함수꼴 $g\left(X_{(n)}\right)$의 기댓값과 $g\left(X_{(n)}\right) = 0$일 확률 사이의 관계를 알아보면 된다.

$$\Pr\left(X_{(n)} \le x\right) = \left(\Pr(X_1 \le x)\right)^n \tag{16}$$
$$= \left(\frac{x}{\theta}\right)^n \tag{17}$$
$$f_{X_{(n)}}(x) = n x^{n-1} \theta^{-n} \tag{18}$$

따라서

$$\begin{aligned}\mathrm{E}\left(g\left(X_{(n)}\right)\right) &= \int_0^\theta n x^{n-1} g\left(x\right) \theta^{-n}\, dx && (19)\\ &= n\theta^{-n}\int_0^\theta x^{n-1} g\left(x\right)\, dx = 0 && (20)\end{aligned}$$

우리가 그러므로 알아봐야 할 관계는 다음과 같다.

$$\int_0^\theta x^{n-1} g\left(x\right)\, dx = 0 \implies g\left(x\right) = 0 \tag{21}$$

미적분의 기본정리에 의해

$$\begin{aligned}\frac{d}{d\theta}\int_0^\theta x^{n-1} g\left(x\right)\, dx &= \theta^{n-1} g\left(\theta\right) && (22)\\ &= 0 && (23)\\ g\left(\theta\right) &= 0 && (24)\end{aligned}$$

0보다 큰 그 어떤 모수값 θ를 가져와도 항상 $g\left(\theta\right) = 0$이므로 $g : \mathbb{R}_+ \mapsto \{0\}$이다.

충분통계량임은 *Neyman-Fisher factorization theorem*을 통해 밝힐 수 있다.

(c) 신뢰구간을 구하기 위해서는 주축통계량(pivot quantity)를 구해야 한다. 주축통계량으로 $X_{(n)}/\theta$를 생각할 수 있다. $Y = X_{(n)}/\theta$라 놓으면

$$f_Y\left(y\right) = ny^{n-1} \tag{25}$$

이 되어 $Y \sim \mathrm{Be}\left(n, 1\right)$이 됨을 알 수 있다. 모수가 n과 1인 베타분포의 CDF를 F라 할 때

$$\Pr\left(F^{-1}_{(1-\alpha_1-\alpha_2)/2} \le \frac{X_{(n)}}{\theta} \le F^{-1}_{(1+\alpha_1+\alpha_2)/2}\right) = 1 - \alpha_1 - \alpha_2 \tag{26}$$

이므로 신뢰구간은

$$F^{-1}_{(1+\alpha_1+\alpha_2)/2} \cdot X_{(n)} \le \theta \le F^{-1}_{(1-\alpha_1-\alpha_2)/2} \cdot X_{(n)} \tag{27}$$

가 될 것이다.

3. (25점) $X_1, \ldots, X_n$이 다음의 확률밀도함수를 가지는 랜덤표본일 때,

$$f\left(x; \theta\right) = \theta x^{\theta-1}, \quad 0 \le x \le 1,\ \theta > 0$$

다음의 가설을 검정하고자 한다.

$$H_0 : \theta = 3 \quad \text{vs} \quad H_1 : \theta \neq 3$$

이 가설에 대한 가능도비 검정을 θ에 대한 최대가능도 추정량(MLE)의 함수로 표현하시오. (즉, MLE의 값에 따라 언제 귀무가설을 기각할 수 있는지 표현하시오.)

Solution: 우선 $X_i \sim \text{Be}\,(\theta, 1)$이다. 2010년 후기 3번 문제에서 구했듯이

$$\begin{aligned}\widehat{\theta}^{\text{MLE}} &= n \Bigg/ \left(-\sum_{i=1}^{n} \ln X_i\right) && (28)\\ &\sim \text{InvGam}\,(n, n\theta) && (29)\end{aligned}$$

이다. 그러므로 가능도비를 구하면

$$\begin{aligned}\frac{L_0}{\widehat{L}} = \left(\frac{3}{\widehat{\theta}}\right)^n \left(\prod_{i=1}^{n} x_i\right)^{3-\widehat{\theta}} &< c_1 && (30)\\ n\left(\ln 3 - \ln \widehat{\theta}\right) + \left(3 - \widehat{\theta}\right)\sum_{i=1}^{n} \ln x_i &< c_2 && (31)\\ -n \ln \widehat{\theta} - \frac{3n}{\widehat{\theta}} &< c_3 && (32)\\ \ln \widehat{\theta} + \frac{3}{\widehat{\theta}} &> c_4 && (33)\end{aligned}$$

도함수와 이계도함수를 구해서 최솟값이 어디인지, 그리고 변곡점이 어디인지 구하면 대충 그래프의 개형이 나온다. $f(x) = \ln x + 3/x$는 $x > 0$인 반직선 위에서 내려갔다 올라가므로 기각역은

$$\text{RR} = \left\{ \widehat{\theta} \,\middle|\, \ln \widehat{\theta} + \frac{3}{\widehat{\theta}} < a \quad \text{or} \quad \ln \widehat{\theta} + \frac{3}{\widehat{\theta}} > b \right\} \tag{34}$$

이고 정확한 a, b값을 알려면 $\ln \widehat{\theta} + 3/\widehat{\theta}$의 분포로 계산을 해야 하는데 변환 자체가 bijective하지 않아서 invertible하지 않으므로 수학적으로 상당히 복잡해진다. 따라서

$$-2 \ln \Lambda \xrightarrow{d} \chi^2_L \tag{35}$$

라는 점근분포(asymptotic distribution)를 이용하는 편이 쉽다.

4. (25점) 선형회귀모형 $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$을 고려하자. 여기서, Y는 독립변수를, $X_1, \ldots, X_p$는 설명변수들을, 그리고 $\beta_0, \ldots, \beta_p$는 회귀계수들을 의미하며 ϵ은 평균이 0, 분산이 σ^2인 오차항을 의미한다.

(a) $p = 1$인 단순선형회귀모형에서 결정계수(coefficient of determination) R^2는 Y와 X_1 사이의 표본상관계수(correlation coefficient)의 제곱과 동일함을 보여라.

(b) 결정계수는 설명변수 $X_1, \ldots, X_p$들의 측정단위에 의존하지 않음을 보여라.

Solution:

(a)

$$\begin{aligned}
R^2 &= 1 - \frac{\text{SSE}}{\text{SST}} && (36)\\
&= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} && (37)\\
&= 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_1\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} && (38)\\
&= 1 - \frac{\sum_{i=1}^{n}\left(y_i - \overline{y} - \widehat{\beta}_1 (x_i - \overline{x})\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} && (39)\\
&= 1 - \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} && (40)\\
&= \frac{\left(\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(x_i - \overline{x})^2} && (41)\\
&= \text{Cor}^2(X, Y) && (42)
\end{aligned}$$

(40)에서 (41)로 넘어갈 때 $\widehat{\beta}_1$을 넣는다.

(b) 설계행렬을 $\mathbf{X}$라 하고 다음과 같이 표기한다.

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{c}_1 & \cdots & \mathbf{c}_p \end{bmatrix} \tag{43}$$

여기서 $\mathbf{c}_i$는 $(i+1)$번째 열을 의미한다. 그렇다면 단위가 다른 설계행렬을 $\mathbf{X}_s$라 하고 다음과 같다고 하자.

$$\mathbf{X}_s = \begin{bmatrix} \mathbf{1} & d_1\mathbf{c}_1 & \cdots & d_p\mathbf{c}_p \end{bmatrix} \tag{44}$$

여기서 d_i는 모두 상수이다. 이는 다시 다음과 같이 표기 할 수 있다.

$$\mathbf{X}_s = \mathbf{X}\mathbf{D} \tag{45}$$

$$\mathbf{D} = \text{diag}(1, d_1, d_2, \ldots, d_p) \tag{46}$$

그리고 Y도 변환해서 aY로 놓는다. 이제 LSE를 다시 구해보면

$$\begin{aligned}\widehat{\beta}_s &= a\mathbf{D}^{-1}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'Y && (47)\\ &= a\mathbf{D}^{-1}\widehat{\beta} && (48)\end{aligned}$$

그리고 적합값(fitted values)도

$$\begin{aligned}\widehat{Y}_s &= \mathbf{X}_s\widehat{\beta}_s && (49)\\ &= a\mathbf{X}\mathbf{D}\mathbf{D}^{-1}\widehat{\beta} && (50)\\ &= a\mathbf{X}\widehat{\beta} && (51)\\ &= a\widehat{Y} && (52)\end{aligned}$$

이제 R^2를 다시 구해보면 다음과 같다. 원래가

$$\mathrm{R}^2 = 1 - \frac{\left(Y-\widehat{Y}\right)'\left(Y-\widehat{Y}\right)}{\left(Y-\overline{y}\mathbf{1}\right)'\left(Y-\overline{y}\mathbf{1}\right)} \tag{53}$$

라면

$$\begin{aligned}\mathrm{R}_s^2 &= 1 - \frac{\left(aY-a\widehat{Y}\right)'\left(aY-a\widehat{Y}\right)}{\left(aY-a\overline{y}\mathbf{1}\right)'\left(aY-a\overline{y}\mathbf{1}\right)} && (54)\\ &= \mathrm{R}^2 && (55)\end{aligned}$$