Daeyoung Lim

Grad School

Year of 2009, late

1. X는 평균 μ_1과 분산 σ^2을 가지는 정규 확률 변수이고 Y는 평균 μ_2와 분산 σ^2을 가지는 정규 확률 변수이며 X와 Y는 서로 독립이라 가정하자. 이때 $U = X + Y$, $V = X - Y$로 정의하자.

   1. U와 V의 결합밀도함수(joint density function)을 구하라.
   2. U와 V가 서로 독립인지 또는 서로 독립이 아닌지 보여라.
   3. U의 적률생성함수(moment generating function)를 구하라.

Solution:

1. U와 V의 자코비언을 구하면 1/2이므로

$$f_{X,Y}(x,y) = \left(2\pi\sigma^2\right)^{-1} \exp\left(-\frac{1}{2\sigma^2}\left((x-\mu_1)^2 + (y-\mu_2)^2\right)\right) \tag{1}$$

$$f_{U,V}(u,v) = \frac{1}{2}\left(2\pi\sigma^2\right)^{-1} \exp\left(-\frac{1}{2\sigma^2}\left(\left(\frac{1}{2}(u+v) - \mu_1\right)^2 + \left(\frac{1}{2}(u-v) - \mu_2\right)^2\right)\right) \tag{2}$$

$$= \left(4\pi\sigma^2\right)^{-1} \exp\left(-\frac{1}{2\sigma^2}\left(\frac{1}{2}\left(u^2+v^2\right) - (\mu_1+\mu_2)u - (\mu_1-\mu_2)v + \mu_1^2 + \mu_2^2\right)\right) \tag{3}$$

2. 결합분포가 separable하므로 독립이다.

$$f_{U,V}(u,v) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\frac{u^2}{2} - (\mu_1+\mu_2)u\right)\right) \exp\left(-\frac{1}{2\sigma^2}\left(\frac{v^2}{2} - (\mu_1-\mu_2)v\right)\right) \tag{4}$$

$$\propto \exp\left(-\frac{1}{4\sigma^2}\left(u - (\mu_1+\mu_2)\right)^2\right) \exp\left(-\frac{1}{4\sigma^2}\left(v - (\mu_1-\mu_2)^2\right)\right) \tag{5}$$

$$U \sim \mathcal{N}\left(\mu_1+\mu_2, 2\sigma^2\right) \tag{6}$$

$$V \sim \mathcal{N}\left(\mu_1-\mu_2, 2\sigma^2\right) \tag{7}$$

   일반적인 정규분포의 특징을 이용하면 동일한 결과가 나온다.

3. 적률생성함수는 $\mathrm{E}\left(e^{tX}\right)$로 정의된다. 하지만 정규분포의 적률생성함수 꼴을 알고 있는 경우 그냥 이용할 수 있다.

$$\mathrm{E}\left(e^{tX}\right) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

   U는 평균이 $\mu_1+\mu_2$고 분산이 $2\sigma^2$이므로 대입하면

$$\mathrm{M}_U(t) = \mathrm{E}\left(e^{tU}\right) = \exp\left((\mu_1+\mu_2)t + \sigma^2 t^2\right).$$

2. 확률변수 $X_1, \ldots, X_n$의 결합확률 밀도함수가 $\prod_{i=1}^{n} f(x_i;\theta)$라고 하자.

   1. 적률방법에 의한 추정량을 설명하고 발생 가능한 문제점을 논하시오.
   2. MLE(maximum likelihood estimator)를 정확히 설명하시오.
   3. MLE의 불변성을 설명하시오.
   4. 정보부등식을 필요한 가정과 더불어 설명하고, 최소분산 추정량(MVUE)를 구하는 데 사용하는 방법을 기술하시오.

Solution:

1. 적률추정법은 모적률과 표본적률을 동치시킴으로써 추정량을 얻는 방법이다. 즉 강대수의 법칙에 의해 표본적률은 모적률로 수렴한다.

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{a.s.} \mathrm{E}\left(X_1^k\right)$$

   그렇기 때문에 표본적률을 통해 얻어낸 추정량은 consistent할 수밖에 없다. 그렇지만 크게 두 가지 문제점이 있다.

   - 추정해야 하는 모수의 개수보다 존재하는 적률의 개수가 더 크면 적률추정량은 유일하지 않다.
   - 적률추정량은 확률변수의 support를 벗어날 수 있다. 예를 들어 $X_1, \ldots, X_n \sim \mathrm{Unif}(0,\theta)$이라면

$$\mathrm{E}(X) = \frac{\theta}{2} = \overline{X}_n$$

     을 통해 얻어낸 $\widehat{\theta} = 2\overline{X}_n$은 얻어낸 표본이 $\theta/2$보다 큰 쪽에 몰려있을 경우 $(0,\theta)$ 구간을 벗어날 수 있다.

2. 최대가능도추정량은 수학적으로는 $\sup L(\theta; X_1, \ldots, X_n)$이 되게 하는 모수값을 의미한다. 직관적으로는 이미 관측된 표본은 이미 관찰된 것이므로 가장 관측되기 쉬워야 한다. 따라서 관측된 표본의 확률값을 가장 높여주는 모수값을 찾는 것이 최대가능도추정량이다. 일반적으로 가장 단순한 분포에서는 support가 모수에 의존하지 않고 모수로 미분된 가능도함수가 모수에 대해 정리될 수 있으면 analytically 표현된다. 하지만

   - support가 모수에 의존할 경우 그래프를 통해 구해야 하며
   - 미분된 가능도함수가 모수에 대해 정리되지 않는 경우 수치적인 방법으로 구해야 한다.

3. 최대가능도추정량은 변환 $T(\theta)$에 대해 불변성을 지닌다는 뜻은 $T\left(\widehat{\theta}\right)$이 최대가능도추정량이 된다는 뜻이다. 만약 변환이 일대응 대응이라면 증명이 쉽지만 일대응 대응 변환이 아닐 경우 *induce likelihood function*을 통해 증명한다.

4. 2008년 후기 2번 문제에도 써놓았으나 다시 한 번 쓰자면 크레이머-라오 정보부등식은 다음 세 가지 가정을 만족해야 한다.

   - 모수 공간이 열린 집합(open set)이거나 닫힌 집합(closed set)일 경우 모수는 그 interior에 속해야 한다.
   - $\mathcal{Y}_\theta = \{y \in \mathcal{Y} \mid f_Y(y \mid \theta) > 0\}$이 모든 $\theta \in \Theta$에 대해 같은 support를 가져야 한다. 즉, 밀도함수가 모수에 의존하지 않아야 한다.
   - 크레이머-라오 정보부등식을 도출하는 과정에서 적분과 미분의 순서를 바꾼다. 그것을 가능하게 해주는 정리가 *Dominated Convergence Theorem*이기 때문에 그 정리의 가정을 만족해야 한다.

   그리하여 나온 정보부등식은 다음과 같다.

$$\operatorname{Var}(T(X)) \geq \frac{\mathrm{E}'(T(X))}{n\mathcal{I}(\theta)}$$

   이것을 MVUE를 구하는 데에 사용하는 법은 통계량의 분산이 정보부등식의 하한과 동일하면 반드시 MVUE이다. 하지만 주의할 것은 모든 MVUE가 전부 하한과 동일한 분산을 가지는 것은 아니다.

3. Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta) = 1/\theta$, where $0 < x < \theta$. We want to test $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Obtain the uniformly most powerful (UMP) test with size α. You must describe how to calculate α to get a full credit.

Solution: Pick θ_1 such that $\theta_0 < \theta_1$. Recall the likelihood function of uniform distribution is constructed as follows:

$$L(\theta \mid x_1, \ldots, x_n) = \frac{1}{\theta^n} I_{(0,\theta)}(x_1) \cdots I_{(0,\theta)}(x_n) \tag{8}$$

$$= \frac{1}{\theta^n} I_{\left(x_{(n)},\infty\right)}(\theta). \tag{9}$$

Then the likelihood ratio becomes

$$\frac{L_1}{L_0} = \left(\frac{\theta_1}{\theta_0}\right)^n \frac{I_{\left(x_{(n)},\infty\right)}(\theta_0)}{I_{\left(x_{(n)},\infty\right)}(\theta_1)}.$$

Since $\theta_0 < \theta_1$, the denominator is implied in the numerator. Thus,

$$\frac{L_1}{L_0} = \left(\frac{\theta_1}{\theta_0}\right)^n I_{(x_{(n)},\infty)}(\theta_0) < k.$$

For this ratio to be small enough so that we could reject the null hypothesis, $X_{(n)}$ should be large enough, almost equal to θ_0 rendering $I_{(x_{(n)},\infty)}(\theta_0) = 0$. Therefore, the rejection region is $X_{(n)} \geq c$ for some constant c. To find the constant, we use the usual identity:

$$\Pr\left(X_{(n)} \geq c \mid H_0\right) = \alpha$$

This is equal to

$$\begin{aligned}
1 - \Pr\left(X_{(n)} < c \mid H_0\right) &= 1 - \Pr\left(X_1 < c \mid H_0\right) \cdots \Pr\left(X_n < c \mid H_0\right) && (10)\\
&= 1 - \left(\Pr\left(X_1 < c \mid H_0\right)\right)^n && (11)\\
&= 1 - \left(\int_0^c \frac{1}{\theta_0}\, dx\right)^n && (12)\\
&= 1 - \left(\frac{c}{\theta_0}\right)^n = \alpha. && (13)
\end{aligned}$$

Therefore, $c = (1-\alpha)^{1/n}\theta_0$. The UMP test for these hypotheses is rejecting the null hypothesis when $X_{(n)} \geq (1-\alpha)^{1/n}\theta_0$.

4. $X_1, \ldots, X_n$이 다음의 확률밀도함수를 갖는 임의표본이라고 하자.

$$f(x \mid \theta) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \;\; 0 < \theta < \infty$$

1. θ의 최대가능도추정량 (MLE) $\hat{\theta}$을 구하라.
2. 다음의 가설검정에 대한 UMP (Uniformly Most Powerful) 기각역의 형태를 구하라.

$$H_0 : \theta = 0.5 \quad \text{vs} \quad H_a : \theta > 0.5$$

3. $n = 10$, $\alpha = 0.05$일 때 2의 UMP 기각역을 χ^2 critical value를 이용하여 나타내어라.

Solution:

1. 로그가능도비를 미분하자.

$$\ln L(\theta; x_1, \ldots, x_n) = n \ln \theta + (\theta - 1) \sum_{i=1}^{n} \ln x_i \tag{14}$$

$$\frac{\partial \ln L(\theta; x_1, \ldots, x_n)}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i \tag{15}$$

$$\widehat{\theta} = \frac{n}{-\sum_{i=1}^{n} \ln x_i} \tag{16}$$

2. 가능도비를 구하면

$$\frac{L_1}{L_0} = \left(\frac{\theta_1}{\theta_0}\right)^n \left(\prod_{i=1}^{n} x_i\right)^{\theta_1 - \theta_0} < k$$

이 되고 이를 정리하면 기각역은

$$\frac{n}{-\sum_{i=1}^{n} \ln X_i} < c \quad \Longrightarrow \quad c' < -\sum_{i=1}^{n} \ln X_i$$

가 된다. 기각역에 있는 통계량의 분포를 알기 위해서는 분포 간의 관계를 이용해야 한다.

- $X_i \sim \mathrm{Be}(\theta, 1)$
- 베타 확률 변수는 범위가 $[0, 1]$이기 때문에 로그를 취하면 $(-\infty, 0]$이 범위다. 따라서 이를 뒤집기 위해서는 마이너스를 붙여야 한다. 이러한 이유로 기각역의 통계량에 마이너스가 붙어있다. 베타 변수에 마이너스 로그를 취하면 지수 분포를 따르는 변수가 된다. 즉 $X \sim \mathrm{Be}(\theta, 1)$이고 $Y = -\ln X$라 하면 자코비언은 $\exp(-y)$이 되므로

  $$f_Y(y) = \theta e^{-y(\theta-1)} \cdot e^{-y} = \theta e^{-\theta y}$$

  따라서 $Y \sim \mathrm{Exp}(\theta)$이다.
- $E_1, \ldots, E_n \overset{iid}{\sim} \mathrm{Exp}(\theta) \equiv \mathrm{Ga}(1, \theta)$이므로

  $$\sum_{i=1}^{n} E_i \sim \mathrm{Ga}(n, \theta)$$

  이다. 고로 $-\sum_{i=1}^{n} \ln X_i \sim \mathrm{Ga}(n, \theta)$이다.
- 감마 확률 변수에 적당한 값을 곱해 카이제곱 변수로 만들어주자.

  $$-2\theta \sum_{i=1}^{n} \ln X_i \sim \mathrm{Ga}\left(n, \frac{1}{2}\right) \equiv \chi^2(2n)$$

상수 c'를 구하기 위해 신뢰수준을 이용하면

$$\Pr\left(-\sum_{i=1}^{10}\ln X_i > c' \,\middle|\, H_0\right) = 0.05$$

이므로 $c' = \chi^2_{20}(0.95)$이다. 여기서 $\chi^2_{20}(0.95)$는 자유도 20인 카이제곱분포의 95% 백분위(percentile)를 뜻한다.

5. A regression analysis (involving 45 observations) relating a dependent variable (Y) and two independent variables resulted in the following information.

$$\hat{y} = 0.408 + 1.3387x_1 + 2x_2 \tag{17}$$

The SSE for the above model is 49.

When two other independent variables were added to the model, the following information was provided.

$$\hat{y} = 1.2 + 3.0x_1 + 12x_2 + 4.0x_3 + 8x_4 \tag{18}$$

This latter model's SSE is 40.

At 95% confidence test to determine if the two added independent variables contribute significantly to the model.

Solution: As explained in the exam of 2009 early, the equation (17) is the model with two restrictions

$$\beta_3 = 0 \tag{19}$$

$$\beta_4 = 0 \tag{20}$$

which translates into a matrix form of

$$\mathbf{A}\beta = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

So the SSE of the model with restrictions, that is equation (17) being the null hypothesis, is

denoted by SSE_H, which is given as 49; whereas, the SSE of the model without any restriction, that is equation (18), is said to be 40. According to the derivation, it is said that the following F-statistic follows an F-distribution with degrees of freedom 2 and $45 - 5$:

$$F = \frac{(\text{SSE}_H - \text{SSE})/2}{\text{SSE}/40} \sim F_{2,40}.$$

Using all the information given, $F = 4.5$. If $4.5 \geq F_{2,40}(0.95)$ where $F_{2,40}(0.95)$ is the 95th percentile of the F-distribution with degrees of freedom 2 and 40, we reject the null hypothesis that the model with restrictions fits better to the data. Otherwise, we conclude that there is little evidence that the restricted model is better.