Daeyoung Lim

Grad School

Year of 2010, early

1.

1. 확률변수 X의 적률생성함수 $M(t)$가 존재한다고 가정할 때, 다음을 보이시오.

$$\begin{cases} \Pr(X \geq a) \leq e^{-ta} M(t), & t > 0 \\ \Pr(X \leq a) \leq e^{-ta} M(t) \end{cases}$$

2. $X_1, \ldots, X_n$이 $\Pr(X_i = 1) = \Pr(X_i = -1) = \frac{1}{2}$를 만족하는 분포에서 나온 임의표본일 때, $S_n = \sum_{i=1}^{n} X_i$에 대하여 다음이 성립함을 1을 이용하여 보이시오.

$$\Pr(S_n \geq a) \leq e^{-ta} e^{nt^2/2}, \quad t > 0$$

3. 어떤 도박꾼이 매 배팅에서 같은 확률로 1원을 따거나 잃는 게임을 하려고 한다고 하자. 매 배팅에서의 결과는 서로 독립이라 할 때, 처음 10번의 배팅에서 적어도 8번 이상 이기는 확률의 상한을 2를 이용하여 구하시오. 또, 이 확률의 정확한 값도 계산하시오.

Solution:

1. 이를 *Chernoff's bound*라고 한다. 마코프 부등식(Markov inequality)로부터 간단하게 도출할 수 있다.

$$\Pr(X \geq a) = \begin{cases} \Pr(tX \geq at) = \Pr\left(e^{tX} \geq e^{at}\right), & t > 0 \\ \Pr(tX \leq at) = \Pr\left(e^{tX} \leq e^{at}\right), & t < 0 \end{cases} \tag{1}$$

$$= \begin{cases} \Pr(X \geq a) \leq e^{-ta} \mathrm{E}\left(e^{tX}\right), & t > 0 \\ \Pr(X \leq a) \leq e^{-ta} \mathrm{E}\left(e^{tX}\right), & t < 0 \end{cases} \tag{2}$$

(1)에서 (2)로 넘어가는 것은 마코프 부등식이다.

2. 이 문제는 표준정규분포를 가정하는 것보다 정확한 분포를 알고 그 분포의 적률생성함수를 이용하는 것이 더 *tight*한 bound를 준다는 것을 증명하는 것이다. 왜냐하면 표준정규확률변수 n개를 더한 변수의 적률생성함수가 $e^{nt^2/2}$이기 때문이다.

$$X \mid 1 \quad -1 \tag{3}$$

$$e^{tX} \mid e^{t} \quad e^{-t} \tag{4}$$

$$\Pr(X = x) \mid \frac{1}{2} \quad \frac{1}{2} \tag{5}$$

이로서 다음과 같은 사실을 얻을 수 있다.

$$\mathrm{E}\left(e^{tX}\right) = \frac{1}{2}\left(e^{t}+e^{-t}\right) \quad \Longrightarrow \quad \mathrm{E}\left(e^{tS_n}\right) = \left(\frac{e^{t}+e^{-t}}{2}\right)^{n} \tag{6}$$

따라서 $2^{-n}\left(e^{t}+e^{-t}\right)^{n} \leq e^{nt^2/2}$임을 보이면 된다.

$$e^{t} = \sum_{k=0}^{\infty}\frac{t^k}{k!} \tag{7}$$

$$e^{-t} = \sum_{k=0}^{\infty}\frac{(-t)^k}{k!} \tag{8}$$

$$e^{t^2/2} = \sum_{k=0}^{\infty}\frac{1}{k!}\left(\frac{t^2}{2}\right)^{k} = \sum_{k=0}^{\infty}\frac{t^{2k}}{k!\times 2^k} \tag{9}$$

(7)과 (8)을 더하면 k가 홀수인 부분은 소거되고 짝수인 부분만 두 배가 된다. 따라서

$$\frac{1}{2}\left(e^{t}+e^{-t}\right) = \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!} \tag{10}$$

분모를 비교하면 $k!\times 2^k \leq (2k)!$이므로 증명끝.

3. 8번 이기면 2번은 진 것이므로 $S_n = 6$이고 9번 이기면 1번은 진 것이므로 $S_n = 8$, 10번 모두 이기면 $S_n = 10$이다. 2번을 이용하면

$$\Pr\left(S_n \geq 6\right) \leq e^{5t^2-6t} \tag{11}$$

이므로 최대값은

$$5\left(t^2 - \frac{6}{5}t + \frac{9}{25}\right) - \frac{9}{5} = 5\left(t-\frac{3}{5}\right)^2 - \frac{9}{5} \tag{12}$$

$t = 3/5$에서 $e^{-9/5}$가 된다. 정확한 확률값은

$$\Pr\left(S_n = 6\right) + \Pr\left(S_n = 8\right) + \Pr\left(S_n = 10\right) = \left(\binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right)\left(\frac{1}{2}\right)^{10} \tag{13}$$

$$= \frac{7}{128} \tag{14}$$

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution on $\left(\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma\right)$. Here the unknown parameters are μ and σ, which are the population mean and standard deviation.

   1. Obtain the probability density function of X_i.
   2. Obtain the likelihood function of μ and σ.

3. Obtain the maximum-likelihood estimators (MLEs) of μ and σ.
4. Obtain the method-of-moments estimators of μ and σ.

Solution:

1. $f_{X_i}(x_i) = \left(2\sqrt{3}\sigma\right)^{-1}$ for $\mu - \sqrt{3}\sigma < x_i < \mu + \sqrt{3}\sigma$

2. 가능도함수의 경우에는 가변수를 통해 support를 명시해 주어야 한다.

$$L(\mu, \sigma; x_1, \ldots, x_n) = \left(2\sqrt{3}\sigma\right)^{-n} \prod_{i=1}^{n} I_{\left(\mu-\sqrt{3}\sigma, \mu+\sqrt{3}\sigma\right)}(x_i) \tag{15}$$
$$= \left(2\sqrt{3}\sigma\right)^{-n} I_{\left(-\infty, x_{(1)}\right)}\left(\mu - \sqrt{3}\sigma\right) I_{\left(x_{(n)}, \infty\right)}\left(\mu + \sqrt{3}\sigma\right) \tag{16}$$

가능도함수 자체는 σ에만 의존하고 단조감소함수이므로 σ가 최대한 작아야 한다. σ의 범위가

$$\frac{1}{\sqrt{3}}\left(\mu - x_{(1)}\right) < \sigma \tag{17}$$
$$\frac{1}{\sqrt{3}}\left(x_{(n)} - \mu\right) < \sigma \tag{18}$$

이기 때문에 부등식의 영역을 통해 σ의 최솟값은 $\mu = 2^{-1}\left(x_{(1)} + x_{(n)}\right)$일 때이고 그 값은

$$\widehat{\sigma} = \frac{1}{2\sqrt{3}}\left(x_{(n)} - x_{(1)}\right) \tag{19}$$

이다. 요약하자면

$$\widehat{\mu} = \frac{1}{2}\left(X_{(1)} + X_{(n)}\right) \tag{20}$$
$$\widehat{\sigma} = \frac{1}{2\sqrt{3}}\left(X_{(n)} - X_{(1)}\right) \tag{21}$$

3. $X \sim \mathrm{Unif}(a, b)$이라 할 때

$$\mathrm{E}(X) = \frac{a+b}{2} \tag{22}$$
$$\mathrm{Var}(X) = \frac{(b-a)^2}{12} \tag{23}$$

이므로

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu \tag{24}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{\left(2\sqrt{3}\sigma\right)^2}{12} + \mu^2 \tag{25}$$

정리하면

$$\widehat{\mu}^{\mathrm{MME}} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{26}$$

$$\widehat{\sigma}^{\mathrm{MME}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2} \tag{27}$$

이고 $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$이다.

3. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with density

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!},$$

where $\lambda \geq 0$ and $x = 0, 1, 2, \ldots$. Let $T = \sum_{i=1}^{n} X_i$.

1. Show that T is complete and sufficient for λ.
2. Show that T/n is the uniformly minimum variance estimator (UMVUE) of λ.
3. Show that $T(T-1)\cdots(T-k+1)/n^k$ is the UMVUE of λ^k for any positive k.

Solution:

1. 포아송 분포에서 표본의 합이 완전통계량(complete statistic)이 됨은 2가지 방법으로 증명할 수 있다.

   - 다음은 완전통계량의 정의를 이용한 방법이다. 통계량 $T(X_1, \ldots, X_n)$에 대해 $f(t;\theta)$가 T의 밀도함수 혹은 질량함수라 할 때 다음이 성립하면 그 분포족을 '완전(complete)'하다 하고, 이와 동치인 명제는 T가 완전통계량이라는 것이다. 모든 모수값 θ에 대해

$$\mathrm{E}\left(r(T)\right) = 0 \quad \Longrightarrow \quad \Pr\left(r(T) = 0\right) = 1.$$

주의해야 하는 것은 '완전성(completeness)'는 하나의 분포족(a family of distributions)의 특성이라는 것이다. 문제의 경우 우리의 통계량은

$$T \sim \text{Poi}(n\lambda)$$

을 따르기 때문에 다음을 보여야 한다.

$$\text{E}(r(T)) = 0 \quad \Longrightarrow \quad \Pr(r(T) = 0) = 1 \tag{28}$$

포아송 분포의 특징 상 지수함수는 음수일 수 없으므로

$$\begin{aligned} \text{E}(r(T)) &= \sum_{t=0}^{\infty} r(t) \frac{e^{-n\lambda}(n\lambda)^t}{t!} && (29) \\ &= \sum_{t=0}^{\infty} r(t) \frac{(n\lambda)^t}{t!} && (30) \\ &= 0 && (31) \end{aligned}$$

모든 λ에 대해 $n\lambda > 0$이므로 모든 $r(t) = 0$이다.

Measure-theoretic하게는 $r(T) \geq 0$이면 적분의 vanishing property에 의해

$$\int r(T)\, dP = 0 \implies r(T) = 0 \;\; P\text{-a.e} \tag{32}$$

이다.

- 두 번째 증명은 지수족에서의 완전통계량 정리를 이용하는 것이다.
  즉, 임의표본 $X_1, \ldots, X_n$가 지수족에 속하는 분포에서 얻은 것일 (모수도 일반적인 상황을 가정해서 k개로 한다.) 때 밀도함수는

$$f(x;\theta) = h(x)\, c(\theta) \exp\left(\sum_{j=1}^{k} \eta_j(\theta)\, t_j(x)\right)$$

이고 결합밀도함수는

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta) = \left(\prod_{i=1}^{n} h(x_i)\right) c(\theta)^n \exp\left(\sum_{j=1}^{k}\left(\eta_j(\theta) \sum_{i=1}^{n} t_j(x_i)\right)\right)$$

이며 이때

$$T(X) = \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

는 완전통계량이다. 포아송분포는 결합밀도함수가

$$f_{X_1,\ldots,X_n}\left(x_1,\ldots,x_n;\lambda\right)=\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)e^{-n\lambda}\exp\left(\ln\lambda\cdot\sum_{i=1}^{n}x_i\right) \tag{33}$$

이기 때문에 $\sum_{i=1}^{n}X_i$이 완전통계량이 된다.

2. *Lehmann-Scheffé* 정리에 의해 T가 완전통계량이자 충분통계량이므로

$$\begin{aligned}\mathrm{E}\left(\frac{T}{n}\right)&=\frac{1}{n}\mathrm{E}\left(T\right) &(34)\\ &=\frac{1}{n}\cdot n\lambda &(35)\\ &=\lambda &(36)\end{aligned}$$

따라서 T/n이 λ의 UMVUE가 된다.

3. 기댓값을 구하면

$$\begin{aligned}\mathrm{E}\left(\frac{T\left(T-1\right)\cdots\left(T-k+1\right)}{n^k}\right)&=\frac{1}{n^k}\sum_{t=0}^{\infty}t\left(t-1\right)\cdots\left(t-k+1\right)\frac{e^{-n\lambda}\left(n\lambda\right)^k}{t!} &(37)\\ &=\frac{1}{n^k}\sum_{t=k}^{\infty}\frac{e^{-n\lambda}\left(n\lambda\right)^{t-k}}{\left(t-k\right)!}\left(n\lambda\right)^k &(38)\\ &=\frac{\left(n\lambda\right)^k}{n^k} &(39)\\ &=\lambda^k &(40)\end{aligned}$$

따라서 증명끝.

4. 세 변수에 대한 n개의 관찰값 (x_{i1},x_{i2},Y_i), $(i=1,\ldots,n)$에 대하여

$$\mathrm{E}\left(Y_i\mid x_{i1},x_{i2}\right)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2},\ \text{그리고 }\mathrm{Var}\left(Y_i\mid x_{i1},x_{i2}\right)=\sigma^2$$

이라 하자.

1. 모수 β_0,β_1,β_2의 최소제곱추정량 $\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\beta}_2$을 유도하시오.
2. $s^2=\dfrac{1}{n-3}\displaystyle\sum_{i=1}^{n}\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1x_{i1}-\widehat{\beta}_2x_{i2}\right)^2$이 σ^2에 대한 불편추정량이 됨을 보이시오.

Solution:

1. 오차항에 정규성 가정은 하지 않고 평균이 0이고 분산이 σ^2을 따른다고 하면 원래 문제의 모형을 오차항을 더해 표현할 수 있다.

$$Y_i \mid x_{i1}, x_{i2} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

이제 행렬표현으로 바꾸자. $Y = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_n \end{bmatrix}'$로 정의하고

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$$

, $\beta = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 \end{bmatrix}'$, $\epsilon = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_n \end{bmatrix}'$라 하자. 그러면 최소제곱법은

$$\frac{d}{d\beta} \left(Y - \mathbf{X}\beta\right)' \left(Y - \mathbf{X}\beta\right) = -2\mathbf{X}'Y + 2\mathbf{X}'\mathbf{X}\beta = 0 \tag{41}$$

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'Y \tag{42}$$

이며 행렬계산은 하지 않는다.

2. 잔차를 자유도로 나눈 것이 오차항의 분산에 대한 비편향추정량이 됨을 증명하는 문제이다.

$$s^2 = \frac{1}{n-3} \left(Y - \mathbf{H}Y\right)' \left(Y - \mathbf{H}Y\right), \qquad \left(\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right) \tag{43}$$

$$= \frac{1}{n-3} Y' \left(\mathbf{I}_n - \mathbf{H}\right) Y, \qquad \left(\left(\mathbf{I}_n - \mathbf{H}\right)^2 = \left(\mathbf{I}_n - \mathbf{H}\right)\right) \tag{44}$$

$$= \frac{1}{n-3} \left(Y - \mathbf{X}\beta\right)' \left(\mathbf{I}_n - \mathbf{H}\right) \left(Y - \mathbf{X}\beta\right), \qquad (\text{전개해보면 원래 부분 제외하고 소거된다}) \tag{45}$$

$$= \frac{1}{n-3} \epsilon' \left(\mathbf{I}_n - \mathbf{H}\right) \epsilon \tag{46}$$

$$\mathrm{E}\left(s^2\right) = \frac{1}{n-3} \operatorname{Tr}\left(\left(\mathbf{I}_n - \mathbf{H}\right) \mathrm{E}\left(\epsilon\epsilon'\right)\right) \tag{47}$$

$$= \frac{1}{n-3} \left(\operatorname{Tr}\left(\mathbf{I}_n\right) - \operatorname{Tr}\left(\mathbf{H}\right)\right) \sigma^2 \tag{48}$$

$$= \frac{\sigma^2}{n-3} \left(\operatorname{Tr}\left(\mathbf{I}_n\right) - \operatorname{Tr}\left(\mathbf{I}_3\right)\right) \tag{49}$$

$$= \sigma^2 \tag{50}$$

5. X_n이 자유도가 n인 카이제곱분포를 따른다고 한다. 이 때

$$Z_n = \frac{X_n - n}{\sqrt{2n}}$$

이라고 하자. n이 무한대로 접근함에 따라 Z_n의 분포가 어느 분포에 수렴하는지를 자세히 보아라. (힌트: 자유도가 n인 카이제곱확률변수는 자유도가 1인 서로 독립인 카이제곱확률변수의 합으로 표현될 수 있다.)

Solution: 자유도가 n인 카이제곱 확률변수는 자유도가 1이고 서로 독립인 카이제곱 확률변수 n개의 합으로 표현할 수 있다. 즉 $Y_1, \cdots, Y_n \sim \chi^2(1)$이라 하면

$$X_n \overset{d}{\equiv} \sum_{i=1}^{n} Y_i.$$

서로 독립인 확률변수의 합으로 표현될 수 있고, 분산이 유한한 확률변수는 중심극한에 의해 정규분포로 수렴한다. 즉

$$\sqrt{n}\frac{\overline{X}_n - \mathrm{E}(Y_1)}{\sqrt{\mathrm{Var}(Y_1)}} \xrightarrow{d} \mathcal{N}(0,1)$$

이므로

$$\begin{aligned} \sqrt{n}\frac{\overline{X}_n - 1}{\sqrt{2}} &= \sqrt{n}\frac{n^{-1}X_n - 1}{\sqrt{2}} && (51)\\ &= \frac{n^{-1/2}X_n - \sqrt{n}}{\sqrt{2}} && (52)\\ &= \frac{X_n - n}{\sqrt{2n}}. && (53) \end{aligned}$$

따라서 Z_n은 표준정규분포로 수렴한다.