استنباط آماری بر روی داده های سانسور شده میانی با توزیع طول عمر نمایی

هنگامه حبیبی راد[۱]، آرزو حبیبی راد و ابوالقاسم بزرگ نیا[۲]

چکیده

در بعضی از مطالعات مربوط به داده های بقا این امکان وجود دارد که بعضی از واحدهای تحت آزمایش، به طور موقتی از مطالعه خارج شده باشند (برای مثال خروج موقتی یک فرد از شهر برای یک مدت نامشخص و برگشت مجدد او به شهر)، در این صورت با سانسور جدیدی به نام سانسور میانی مواجه می شویم که اولین بار توسط جامالامادا کا و مانگالام[۳][٤] معرفی شده است.

فرض کنید n واحد مشابه با زمان های بقای $T_1,T_2,...,T_n$ تحت یک آزمایش بقا قرار گرفته باشند و برای iامین واحد یک بازه سانسور تصادفی (L_i,R_i) وجود داشته باشد، به طوری که زمان T_i برای iامین واحد در صورتی ثبت می شود که $T_i \notin [L_i,R_i]$ و در غیر این صورت سانسور می شود. یعنی اگر طول عمر یک واحد تحت آزمایش، در یک بازه تصادفی قرار گیرد، سانسور میانی رخ می دهد.

به دنبال ارائه این نوع سانسور آیر[٤] و همکاران [۳] تحلیل داده های سانسور شده میانی را با فرض نمایی بودن توزیع طول عمر مورد بررسی قرار داده اند. در این مقاله پس از معرفی کامل طرح سانسور میانی، ابتدا برآورد درستنمایی ماکسیمم آن را بررسی می کنیم و سپس به بیان نتایج تئوری و قضایای مربوطه می پردازیم.

واژه های کلیدی: برآورد درستنمایی ماکسیمم، توزیع طول عمر، توزیع نمایی، سانسور میانی.

۱– مقدمه

در این مقاله با داده هایی سروکار داریم که دارای سانسور میانی هستند، سانسور میانی زمانی رخ می دهد که داده در فاصله تصادفی قرار بگیرد و مقدار واقعی آن، مشاهده نشود. طرح سانسور میانی به صورت زیر تعریف می شود:

فرض کنید n واحد یکسان مورد آزمایش قرار می گیرند و طول عمر این واحدها، $T_1,T_2,...,T_n$ هستند. برای iامین واحد، T_i فقط زمانی مشاهده می شود که $T_i \notin [L_i,R_i]$، در غیر این صورت مشاهده نمی شود. فرض کنید $\delta_i = I(T_i \notin [L_i,R_i])$، که در آن $I(\cdot)$ تابع نشانگر در نظر گرفته شده است، یعنی وقتیکه $\delta_i = 1$، مشاهدات سانسور نشده اند و مقدار واقعی T_i مشاهده می شود، در این شرایط ما (L_i,R_i) را مشاهده نمی کنیم. در شرایط دیگر، زمانیکه $\delta_i = 0$، فقط فاصله سانسور شده $[L_i,R_i]$ را مشاهده می کنیم، در این صورت مشاهده سانسور شده است. برای iامین واحد داریم:

$$(Y_i,\delta_i) = \begin{cases} (T_i,1), & T_i \notin [L_i,R_i], \\ ([L_i,R_i],0), & o.w. \end{cases} \qquad (1)$$

بنابراین، داده هایی که در حالت سانسور نشده بدست آمده اند با داده هایی که در شرایط فاصله سانسور شده به دست آمده اند یکسان نیستند. بر اساس مشاهدات، بدست آوردن تابع توزیع طول عمر T_i و گسترش روش های استنتاجی مورد نیاز مشکل به نظر می رسد.

در ابتدا به بیان چند مثال که در آن ها ممکن است سانسور میانی رخ دهد می پردازیم. فرض کنید در یک مطالعه طول عمر، موضوع مورد بررسی به طور موقت از مطالعه خارج

---

[1] دانشجوی کارشناسی ارشد دانشگاه آزاد اسلامی مشهد

[2] عضو هیئت علمی گروه آمار، دانشگاه فردوسی مشهد و دانشگاه آزاد اسلامی مشهد

[3] Jammalamadaka and Mangalam

[4] Iyer

شود (مثلاً، اگر فردی به طور موقت در یک دوره زمانی شـهر را ترک کند و برگردد، در حالیکه هنوز زنده اسـت) سانـسور میانی رخ می دهد. همچنینی سانسور میانی زمانی رخ می دهد که مثلاً وسیله اندازه گیری در یک دوره زمانی خراب شود یـا اگر یک کلینیک که در آن مشاهدات جمع آوری می شـود در یک دوره به دلیل شرایط ناگهانی مانند وقوع یک جنگ یا یک حادثه بسته شود. در چنین شرایطی، رخداد مورد نظر در مدت زمانی که یک مورد رخ نداده و یا مشاهده نمی شود، اتفاق می افتد.

در مقاله جامالاماداکـا و مانگـالام[٤] $T_1,T_2,...,T_n$ متغیرهـای تصادفی مستقل و هم توزیع با تابع توزیـع نـامعلوم $F(\cdot)$ در نظر گرفته شده اند، و $(L_1,R_1),...,(L_n,R_n)$ مستقل و هم توزیـع هستند با تابع توزیع دو متغیره نامعلوم $G(\cdot,\cdot)$ و مستقل از T_i. البته در مقاله ذکر شده مـسئله بـه صـورت ناپـارامتری مـورد بررسی قـرار گرفتـه اسـت ولـی در ایـن مقالـه فرمـول هـای پارامتری مورد بررسی قرار می گیرند. برای این منظور فـرض کنید که $T_1,T_2,...,T_n$ متغیرهای تصادفی مستقل و هم توزیـع نمایی با میانگین مثلاً $\theta_0=1/2$ با تابع چگالی (pdf) زیر باشد:

$$f(x;\theta_0)=\begin{cases}\theta_0 e^{-\theta_0 x} & , \quad x>0,\\ 0 & , \quad o.w,\end{cases} \qquad (2)$$

علاوه بر این، $(L_1,Z_1),...,(L_n,Z_n)$ نیـز مـستقل و هـم توزیـع هستند که L_i و $Z_i=R_i-L_i$ متغیرهای تصادفی مستقل نمـایی و مستقل از T_i هستند. هم چنین فرض می شود کـه L_i و Z_i به ترتیب دارای میـانگین هـای $1/\alpha$ و $1/\beta$ اسـت و بـه θ_0 بستگی ندارد. البته باید در نظر داشت فراینـد سانـسور شـدن، مستقل از طول عمر جامعه مورد نظر است و هـیچ اطلاعـاتی در مورد طول عمر نمی دهد.

فرض هایی که در مورد استقلال، در تحلیـل داده هـای طـول عمر می شود کاملاً استاندارد هستند. برای مثـال بـه کـاپلان و میر[1][٦]، بابو و همکاران[2][١]، جامالاماداکا و مانگالام[٤]، چـو و تیچر[3][٢]، جامالاماداکا و آیر[4][٥] و مراجع ذکر شـده در آن، که فرض های متفاوتی برای طرح های سانسور ساخته شـده، مراجعه کنید. به طور کلی دلایل مختلفی برای استقلال وجـود دارد که به بعضی از آن ها اشاره می کنیم. در ابتـدا، در بیـشتر شرایطی که در واقعیت رخ می دهـد احتمـال بـستگی داشـتن فرایند سانسور شدن به طول عمر جامعه مورد نظر کم است و این مطلب اطلاعاتی در مورد تابع توزیع جامعه به ما می دهد. در مثال هایی که قبلاً بیان کـردیم، سانـسور میـانی بـه دلایـل خارجی رخ می دهد، که هیچ ارتبـاطی بـا متغیـر طـول عمـر ندارد.

در ادامه مقاله، ابتدا در بخـش دوم دو روش تکـرار عـددی و الگوریتم EM را برای برآورد پارامتر مجهول تابع درسـتنمایی ماکسیمم معرفی کرده و سپس در بخش سوم بـا کمـک چنـد قضیه و چند لم خصوصیات مربوط به تابع درستنمایی معرفی شده در بخش دوم را، ثابـت مـی کنـیم و در ادامـه در بخـش چهارم نتایج گرفته شده از این مقاله بیان شده است.

۲– برآورد درستنمایی ماکسیمم برای توزیع طول عمر نمایی

ابتدا داده ها را با فرض نداشتن داده گم شـده مرتـب مـی کنیم، به این صورت که n_1 جمله اول را مـشاهدات سانـسور نشده و n_2 جمله بعدی را مشاهدات سانـسور شـده در نظـر می گیریم. بنابراین مشاهدات به صورت زیر خواهد بود:

$$(T_1,1),...,(T_{n_1},1),\ (L_{n_1+1},R_{n_1+1}),...,(L_{n_1+n_2},R_{n_1+n_2}), \qquad (3)$$

کـــه در آن $n_1+n_2=n$. بنـــابراین، بـــرای n_1 مـــشاهده اول $T_i\in[L_i,R_i]$ و برای n_2 مشاهده بعدی $T_i\notin[L_i,R_i]$.

لم۲-۱: تابع درستنمایی داده هـای سانـسور شـده میـانی بـه صورت زیر است

$$l(\theta)=c\,\theta^{n_1}e^{-\theta\sum_{i=1}^{n_1}t_i}\prod_{i=n_1+1}^{n_1+n_2}\left(e^{-\theta l_i}-e^{-\theta r_i}\right), \qquad (4)$$

که در آن c یک مقدار ثابت بوده و تنها بـه α و β بـستگی دارد. چون علاقه ای به برآورد α و β نـداریم، در مـورد آن بحث نمی کنیم.

اثبات: با توجه به تعریف سانسور میانی داریم

$$l(\theta)=\prod_{i=1}^{n_1}f_{T_i,\delta}(t_i,1)\prod_{i=n_1+1}^{n_1+n_2}f_{L_i,Z_i,\delta}(l_i,z_i,0),$$

از طرفی

[1] Kaplan and Meier
[2] Babu, Rao and Rao
[3] Chow and Teicher
[4] Jammalamadaka and Iyer

$$f_{T_i,\delta}(t_i,1) = P(\delta=1|T_i=t_i) f_{T_i}(t_i), \quad (5)$$

که در آن $f_{T_i}(t_i)=\theta e^{-\theta t_i}$ و

$$P(\delta=1|T_i=t_i) = P(T_i \notin (L_i,R_i)|T_i=t_i) = 1-\frac{\alpha e^{-\beta t_i}}{\alpha-\beta}\left(1-e^{-(\alpha-\beta)r_i}\right),$$

همچنین

$$f_{L_i,Z_i,\delta}(l_i,z_i,0) = P(\delta=0|L_i=l_i,Z_i=z_i) = e^{-\theta l_i}-e^{-\theta r_i}. \quad (6)$$

در انتها با جایگذاری روابط (۵) و (٦) در فرمول (٤) لم فوق ثابت می شود.

در ادامه به دنبال محاسبه MLE برای پارامتر θ هستیم، که بر اساس فرمول (٤) لگاریتم درستنمایی

به صورت زیر می باشد:

$$\ln l(\theta) = L(\theta) = \ln c + n_1 \ln\theta - \theta\sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} \ln\left(e^{-\theta l_i}-e^{-\theta r_i}\right). \quad (7)$$

با مشتق گیری از $L(\theta)$ و برابر صفر قرار دادن، داریم

$$\frac{\partial L(\theta)}{\partial\theta} = \frac{n_1}{\theta} - \sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} \frac{(r_i-l_i)}{e^{\theta(r_i-l_i)}-1} - \sum_{i=n_1+1}^{n_1+n_2} l_i = 0. \quad (8)$$

بنابراین $\hat{\theta}$ که از حل معادله (۸) حاصل می شود، یک MLE برای پارامتر θ است. ولی با استفاده از معادله (۸) نمی توان $\hat{\theta}$ را مستقیماً و به طور دقیق بدست آورد، لذا برای بدست آوردن MLE از روش های تکراری استفاده می کنیم. توجه کنید که برای این منظور رابطه (۸) را می توان به صورت زیر نوشت

$$h(\theta) = \theta, \quad (9)$$

که با استفاده از تعریف $z_i = r_i - l_i$ و با کمک رابطه (۸) واضح است که

$$h(\theta) = \frac{1}{\sum_{i=n_1+1}^{n_1+n_2} l_i + \sum_{i=1}^{n_1} t_i}\left[n_1 + \theta\sum_{i=n_1+1}^{n_1+n_2}\frac{z_i e^{-\theta z_i}}{1-e^{-\theta z_i}}\right].$$

بنابراین یک روش ساده برای بدست آوردن MLE استفاده از روش تکرار برای معادله (۹) است. برای مثال می توانیم با مقدار اولیه $\theta^{(1)}$ شروع کنیم، سپس $\theta^{(2)}=h(\theta^{(1)})$ بدست می آید و به همین ترتیب ادامه می دهیم. در این روش تکراری زمانی متوقف می شویم که $|\theta^{(i)}-\theta^{(i+1)}|<\varepsilon$، و در آن ε مقدار کوچک مثبتی است که از قبل تعیین می شود. برای انتخاب اول θ می توان از $\theta^{(1)}=\frac{n_1}{\sum_{i=1}^{n_1} t_i}$ استفاده کرد. یکی از روش های دیگر برای بدست آوردن MLE استفاده از الگوریتم EM است، که در این روش ابتدا مقدار $E(T|L<T<R)$ را بدست می آوریم، که در آن L و R مقادیر ثابت هستند و T دارای توزیع نمایی با میانگین $1/\theta$ است،

$$E(T|L<T<R) = \int_L^R t f_{T|T\in(L,R)}(t|L<t<R)dt,$$

برای محاسبه رابطه فوق ابتدا لم زیر را ثابت می کنیم.

لم۲۰۲: تابع چگالی T، به شرط آنکه $T\in(L,R)$ به صورت زیر است

$$f_{T|T\in(L,R)}(t|L<t<R) = \frac{\theta e^{-\theta t}}{e^{-\theta L}-e^{-\theta R}}.$$

اثبات: می دانیم

$$f_{T|T\in(L,R)}(t|L<t<R) = \frac{P(T=t,T\in(L,R))}{P(T\in(L,R))} = \frac{f_T(t)}{P(T\in(L,R))},$$

که در آن $f_T(t)=\theta e^{-\theta t}$ و

$$P(T\in(L,R)) = P(L<T<R) = e^{-\theta L}-e^{-\theta R}.$$

در نتیجه لم ثابت می شود.

با کمک لم فوق داریم

$$E(T|L<T<R) = \int_L^R t f_{T|T\in(L,R)}(t|L<t<R)dt = \int_L^R \frac{t\theta e^{-\theta t}}{e^{-\theta L}-e^{-\theta R}}dt,$$

پس

$$E(T|L<T<R) = \frac{e^{-\theta L}\left(L+\frac{1}{\theta}\right)}{e^{-\theta L}-e^{-\theta R}} - \frac{e^{-\theta R}\left(R+\frac{1}{\theta}\right)}{e^{-\theta L}-e^{-\theta R}}. \quad (10)$$

توجه کنید که فرمول (۱۰) در محاسبه الگوریتم EM (برای آشنایی بیشتر با این الگوریتم میتوانید به کتاب :Simulation Modeler's Approach نوشته شده توسط جیمز تامپسون[۷] مراجعه کنید) استفاده می شود. در این صورت تابع شبه درستنمایی به شکل زیر در می آید:

$$l(\theta)=\theta^{n_1+n_2}e^{-\theta\left(\sum_{i=1}^{n_1}T_i+\sum_{i=n_1+1}^{n_1+n_2}T_i^{(s)}\right)}, \qquad (۱۱)$$

که در آن

$$T_i^{(s)}=\frac{e^{-\theta L_i}\left(L_i+\frac{1}{\theta}\right)-e^{-\theta R_i}\left(R_i+\frac{1}{\theta}\right)}{e^{-\theta L_i}-e^{-\theta R_i}}. \qquad (۱۲)$$

مراحل الگوریتم EM عبارتند از:

مرحله۱: فرض می کنیم θ_j، j امین تکرار $\hat{\theta}$ باشد.

مرحله۲: $T_{i(j)}^{(s)}$ را با استفاده از فرمول (۱۲) محاسبه می کنیم و θ را توسط $\theta_{(j)}$ جایگزین می کنیم.

مرحله۳: $\theta_{(j+1)}=\frac{n_1+n_2}{\sum_{i=1}^{n_1}T_i+\sum_{i=n_1+1}^{n_1+n_2}T_{i(j)}^{(s)}}$.

۳- نتایج نظری

قضیه ۳-۱: روش های تکراری مطرح شده همگرا هستند اگر

$$\sum_{i=n_1+1}^{n_1+n_2}r_i\le 2\sum_{i=1}^{n_1}t_i+3\sum_{i=n_1+1}^{n_1+n_2}l_i. \qquad (۱۳)$$

اثبات: با مشتق گیری از $h(\theta)$ داریم

$$\frac{\partial h(\theta)}{\partial\theta}=h'(\theta)=\frac{1}{\sum_{i=n_1+1}^{n_1+n_2}l_i+\sum_{i=1}^{n_1}t_i}\left[\sum_{i=n_1+1}^{n_1+n_2}\frac{z_ie^{-\theta z_i}}{1-e^{-\theta z_i}}+\theta\sum_{i=n_1+1}^{n_1+n_2}\frac{-z_i^2e^{-\theta z_i}\left(1-e^{-\theta z_i}\right)-\left(z_ie^{-\theta z_i}\right)^2}{\left(1-e^{-\theta z_i}\right)^2}\right],$$

پس

$$|h'(\theta)|=\frac{1}{\sum_{i=n_1+1}^{n_1+n_2}l_i+\sum_{i=1}^{n_1}t_i}\times\left|\sum_{i=n_1+1}^{n_1+n_2}\frac{z_ie^{-\theta z_i}\left(1-e^{-\theta z_i}-\theta z_i\right)}{\left(1-e^{-\theta z_i}\right)^2}\right|$$

به سادگی می توان نشان داد که برای هر $x\ge 0$

$$\frac{|e^{-x}||(1-e^{-x}-x)|}{|1-e^{-x}|^2}\le\frac{1}{2},$$

بنابراین

$$|h'(\theta)|\le\frac{1}{2}\frac{\sum_{i=n_1+1}^{n_1+n_2}z_i}{\sum_{i=n_1+1}^{n_1+n_2}l_i+\sum_{i=1}^{n_1}t_i}.$$

از طرفی می دانیم اگر روش های تکراری همگرا باشند $|h'(\theta)|\le 1$ (اسماعیلی[۹])، در این صورت رابطه (۱۳) برقرار می باشد.

لم ۱۰۳: توجه کنید که بنابر رابطه

$$\frac{1}{n}L(\theta)=c'+\frac{n_1}{n}\ln\theta-\frac{\theta}{n}\sum_{i=1}^{n_1}t_i-\frac{\theta}{n}\sum_{i=n_1+1}^{n_1+n_2}l_i+\frac{1}{n}\sum_{i=n_1+1}^{n_1+n_2}\ln\left(1-e^{-\theta z_i}\right),$$

به سادگی داریم

$$\frac{1}{n}L(\theta)\to g(\theta)\quad a.s.$$

که در آن

$$g(\theta)=c'+p(\theta_0)\ln\theta+\theta\frac{(1-p(\theta_0))(\alpha+\beta+2\theta_0)}{(\alpha+\theta_0)(\beta+\theta_0)}-\theta\frac{1}{\theta_0}-\theta\frac{(1-p(\theta_0))}{\alpha+\theta_0}-\frac{\alpha\beta}{\alpha+\theta_0}\left[\sum_{i=1}^{\infty}\frac{1}{i(\beta+i\theta)}-\sum_{i=1}^{\infty}\frac{1}{i(\beta+i\theta+\theta_0)}\right],$$

و

$$p(\theta)=\frac{\alpha\beta+\beta\theta+\theta^2}{(\alpha+\theta)(\beta+\theta)},\qquad c'=\frac{1}{n}\ln c, \qquad (۱۴)$$

اثبات: ابتدا تابع چگـالی T را بـه شـرط آن کـه $T\notin(L,R)$ و $\alpha\neq\beta$، به صورت زیر محاسبه می کنیم. با فرض

$$\delta=\begin{cases}1 & T\notin(L,R)\\ 0 & T\in(L,R),\end{cases}$$

داریم

$$f_{T|\delta}(t|1)=\frac{f_{T,\delta}(t,1)}{P(\delta=1)}=\frac{P(\delta=1|T=t)f_T(t)}{p(\theta_0)},$$

که در آن

$$P(\delta=1|T=t)=P(T\notin(L,R)|T=t)=1-P(L<t<R).$$

از طرفی می دانیم $R=L+Z$، پس

$$P(L<t<R)=\int_0^t\int_{t-l}^{\infty}f(l,z)dzdl=\frac{\alpha e^{-\beta t}}{\alpha-\beta}\left(1-e^{-(\alpha-\beta)t}\right),$$

و

$$P(\delta=1|T=t)=1-\frac{\alpha e^{-\beta t}}{\alpha-\beta}\left(1-e^{-(\alpha-\beta)t}\right),$$

در ادامه نشان می دهیم کـه $p(\theta)=\frac{\alpha\beta+\beta\theta+\theta^2}{(\alpha+\theta)(\beta+\theta)}$، بـرای این منظور داریم

$$P_\theta(T\notin(L,R))=1-P_\theta(T\in(L,R))=1-\left[P_\theta(T<L+Z)-P_\theta(T<L)\right],$$

از طرفی

$$P_\theta(T<L+Z)=\int_0^\infty\int_0^\infty\int_0^{l+z}(\alpha e^{-\alpha l}\beta e^{-\beta z}\times\theta e^{-\theta t})dtdldz$$
$$=\frac{\alpha\theta+\beta\theta+\theta^2}{(\alpha+\theta)(\beta+\theta)},$$

و

$$P_\theta(T<L)=\int_0^\infty\int_0^{l}(\alpha e^{-\alpha l}\theta e^{-\theta t})dtdl$$
$$=\frac{\theta}{(\alpha+\theta)},$$

در نتیجه

$$p(\theta)=P_\theta(T\notin(L,R))=\frac{\alpha\beta+\beta\theta+\theta^2}{(\alpha+\theta)(\beta+\theta)}.$$

پس با فرض $\theta=\theta_0$ داریم

$$f_{T|T\notin(L,R)}(t)=\frac{1}{p(\theta_0)}\left\{\theta_0 e^{-\theta_0 t}\left(1-\frac{\alpha e^{-\beta t}}{\alpha-\beta}\left(1-e^{-(\alpha-\beta)t}\right)\right)\right\}. \quad (15)$$

و به همین ترتیب در حالت $\alpha=\beta$، داریم

$$P(L<t<L+Z)=\int_0^t\int_{t-l}^\infty f(l,z)dzdl$$
$$=\alpha te^{-\alpha t},$$

و

$$P(\delta=1|T=t)=1-\alpha te^{-\alpha t},$$

در نتیجه

$$f_{T|T\notin(L,R)}(t)=\frac{1}{p(\theta_0)}\{\theta_0 e^{-\theta_0 t}(1-\alpha te^{-\alpha t})\}. \quad (16)$$

با کمک فرمول های (۱۵) و (۱۶) داریم

$$E(T|T\notin(L,R))=\int_0^\infty t f_{T|T\notin(L,R)}(t)dt$$
$$=\begin{cases}\frac{1}{p(\theta_0)}\left[\frac{1}{\theta_0}-\frac{2\alpha\theta_0}{(\alpha+\theta_0)^2}\right], & \alpha=\beta\\ \frac{1}{p(\theta_0)}\left[\frac{1}{\theta_0}-\frac{\alpha\theta_0}{(\alpha-\beta)}\left(\frac{1}{(\beta+\theta_0)^2}-\frac{1}{(\alpha+\theta_0)^2}\right)\right], & \alpha\neq\beta\end{cases}$$

در ادامه تابع چگالی L را به شـرط آن کـه $T\in(L,R)$ باشـد محاسبه می کنیم،

$$f_{L|T\in(L,R)}(l)=\frac{P(L=l,T\in(L,R))}{P(T\in(L,R))},$$
$$=\frac{P(T\in(L,R)|L=l)f_L(l)}{P(T\in(L,R))},$$

که با فرض $\theta=\theta_0$ داریم

$$P_\theta(T\in(L,R))=1-p(\theta_0),$$

و $f_L(l)=\alpha e^{-\alpha l}$. هم چنین

$$P(T\in(L,R)|L=l)=P(l<T<l+Z)$$
$$=P(T-Z<l)-P(T<l),$$

$$P(T-Z<l)=\int_0^\infty\int_0^{l+z}(\beta e^{-\beta z}\theta_0 e^{-\theta_0 t})dtdz$$
$$=1-\frac{\beta e^{-\theta_0 l}}{(\beta+\theta_0)},$$

و

$$P(T<l)=\int_0^l\theta_0 e^{-\theta_0 t}dt=1-e^{-\theta_0 l},$$

در نتیجه برای $l>0$ داریم

$$f_{L|T\in(L,R)}(l)=\frac{1}{1-p(\theta_0)}\times\frac{\alpha\theta_0}{(\beta+\theta_0)}e^{-(\alpha+\theta_0)l}.$$

در این صورت

$$E(L|T\in(L,R))=\int_0^\infty l f_{L|T\in(L,R)}(l)dl$$
$$=\frac{1}{1-p(\theta_0)}\times\frac{\alpha\theta_0}{(\beta+\theta_0)(\alpha+\theta_0)^2}.$$

همچنین تابع چگالی $Z=R-L$ را به شرط آن کـه $T\in(L,R)$ به صورت زیر بدست می آوریم

$$f_{Z|T\in(L,R)}(z)=\frac{P(Z=z,T\in(L,R))}{P(T\in(L,R))},$$
$$=\frac{P(T\in(L,R)|Z=z)f_Z(z)}{P(T\in(L,R))},$$

که در آن $f_Z(z)=\beta e^{-\beta z}$ و

$$P_\theta(T\in(L,R))=1-p(\theta_0),$$

و

$$P(T\in(L,R)|Z=z)=P(L<T<L+z)$$
$$=P(T-L<z)-P(T<L),$$

که در رابطه فوق

$$P(T-L<z)=\int_0^\infty\int_0^{l+z}(\alpha e^{-\alpha l}\theta_0 e^{-\theta_0 t})dtdl$$
$$=1-\frac{\alpha e^{-\theta_0 z}}{(\alpha+\theta_0)},$$

و

$$P(T<L)=\int_0^\infty\int_0^{l}\alpha e^{-\alpha l}\theta_0 e^{-\theta_0 t}dtdl$$
$$=\frac{\theta_0}{(\alpha+\theta_0)}.$$

در نتیجه برای $z>0$ داریم

$$f_{Z|T\in(L,R)}(z)=\frac{1}{1-p(\theta_0)}\times\frac{\alpha\beta e^{-\beta z}}{(\alpha+\theta_0)}(1-e^{-\theta_0 z}).$$

از طرفی امید تابع $\ln(1-e^{-\theta z})$ به شرط $T\in(L,R)$ بـه صـورت زیر محاسبه می شود

$$E\left(\ln\left(1-e^{-\theta z}\right)\middle|T\in(L,R)\right)=\int_0^\infty \ln\left(1-e^{-\theta z}\right)\times f_{Z|T\in(L,R)}(z)\,dz$$
$$=\frac{1}{1-p(\theta_0)}\times\frac{\alpha\beta}{(\alpha+\theta_0)}\times\left[\int_0^\infty e^{-\beta z}\ln\left(1-e^{-\theta z}\right)dz-\int_0^\infty e^{-(\beta+\theta_0)z}\ln\left(1-e^{-\theta z}\right)dz\right].$$

برای محاسبه مقـدار انتگـرال بـالا از بـسط $\ln(1-e^{-\theta z})$ (عـالم زاده[۱۰]) استفاده می کنیم

$$\ln\left(1-e^{-\theta z}\right)=\sum_{i=1}^{\infty}\frac{-e^{-i\theta z}}{i},\qquad \left|e^{-\theta z}\right|<1$$

در نتیجه

$$\int_0^\infty e^{-\beta z}\ln\left(1-e^{-\theta z}\right)dz=\int_0^\infty e^{-\beta z}\sum_{i=1}^{\infty}\frac{-e^{-i\theta z}}{i}dz=-\sum_{i=1}^{\infty}\frac{1}{i(\beta+i\theta)},$$

و

$$\int_0^\infty e^{-(\beta+\theta_0)z}\ln\left(1-e^{-\theta z}\right)dz=\int_0^\infty e^{-(\beta+\theta_0)z}\times\sum_{i=1}^{\infty}\frac{-e^{-i\theta z}}{i}dz=-\sum_{i=1}^{\infty}\frac{1}{i(\beta+i\theta+\theta_0)}.$$

پس داریم

$$E\left(\ln\left(1-e^{-\theta z}\right)\middle|T\in(L,R)\right)=-\frac{1}{1-p(\theta_0)}\times\frac{\alpha\beta}{(\alpha+\theta_0)}\times\left[\sum_{i=1}^{\infty}\frac{1}{i(\beta+i\theta)}-\sum_{i=1}^{\infty}\frac{1}{i(\beta+i\theta+\theta_0)}\right].$$

در ادامـه بـر اسـاس قـانون قـوی اعـداد بـزرگ بـا فـرض

$$S_n=\sum_{i=1}^{n}X_i=n_1$$

$$\frac{n_1}{n}\to p(\theta_0)\quad a.s.$$

و با فرض $n_2=n-n_1$

$$\frac{n_2}{n}\to 1-p(\theta_0)\quad a.s.$$

همچنین

$$\frac{\sum_{i=1}^{n_1}T_i}{n_1}\to E\left(T_i\middle|T_i\notin(L_i,R_i)\right)\quad a.s.$$

و

$$\frac{\sum_{i=n_1+1}^{n_1+n_2}L_i}{n_2}\to E\left(L_i\middle|T_i\in(L_i,R_i)\right)\ a.s.$$

و

$$\frac{\sum_{i=n_1+1}^{n_1+n_2}\ln\left(1-e^{-\theta Z_i}\right)}{n_2}\to E\left(\ln\left(1-e^{-\theta Z_i}\right)\middle|T_i\in(L_i,R_i)\right)\ a.s.$$

پس می توان به راحتی نشان داد که $\frac{1}{n}L(\theta)\to g(\theta)\ \ a.s.$.

لم ۳-۲: $g(\theta)$ یک تابع تک نمایی با ماکسیمم واحد است.

اثبات: با توجه به تعریف تابع $g(\theta)$ در رابطه (۱٤) داریم

$$g'(\theta)=\frac{p(\theta_0)}{\theta}+\frac{(1-p(\theta_0))(\alpha+\beta+2\theta_0)}{(\alpha+\theta_0)(\beta+\theta_0)}-\frac{1}{\theta_0}-\frac{(1-p(\theta_0))}{\alpha+\theta_0}-\frac{\alpha\beta}{\alpha+\theta_0}\left[-\sum_{i=1}^{\infty}\frac{1}{(\beta+i\theta)^2}+\sum_{i=1}^{\infty}\frac{1}{(\beta+i\theta+\theta_0)^2}\right],$$

در نتیجه

$$g'(0)=\infty$$

و

$$g'(\infty)=-\frac{\beta+\theta_0 p(\theta_0)}{\theta_0(\beta+\theta_0)}<0.$$

و

$$g''(\theta)=-\frac{p(\theta_0)}{\theta^2}-\frac{\alpha\beta}{\alpha+\theta_0}\times\left[\sum_{i=1}^{\infty}\frac{2i}{(\beta+i\theta)^3}+\sum_{i=1}^{\infty}\frac{2i}{(\beta+i\theta+\theta_0)^3}\right]<0.$$

روابط فوق نشان می دهند کـه $g(\theta)$ فقـط دارای یـک ریـشه است و چون مشتق دوم همواره منفی است پس آن ریشه یک نقطه ماکسیمم است. پس $g(\theta)$ یک تابع تک نمـایی بـا یـک نقطه ماکسیمم است.

لم ۳-۳: اگر $\hat{\theta}$ برآورد درستنمایی θ_0 باشـد، آنگـاه $\hat{\theta}\to\theta^*$ که در آن θ^* جواب یکتای معادله غیر خطی زیر است

$$g'(\theta)=0 \qquad (۱۷)$$

که $p(\theta)$ در (۱٤) تعریف شده است.

اثبات: بنابر لم ۲، رابطه (۱۷) تنها یک جـواب ماننـد θ^* دارد زیرا $g'(\theta)$ محور xها را تنها در یک نقطه قطـع مـی کنـد و

$g''(\theta)<0$ به ازای هـر θ. اثبـات را در حالـت اول: $\hat{\theta}_n$ بـرای همه nها کراندار است، و در حالت دوم: $\hat{\theta}_n$ کرانـدار نیـست ادامه داده و در آن $\hat{\theta}$ را با $\hat{\theta}_n$ نشان می دهیم.

حالت۱: $\hat{\theta}_n$ برای همه nها کراندار است.

فرض کنید $\hat{\theta}_n$ به θ^* همگرا نباشد. بنابراین، زیر مجموعـه ای ماننـد $\{n_k\}$ از $\{n\}$ وجـود دارد کـه بـرای $\tilde{\theta} \neq \theta^*$، داریـم $\hat{\theta}_{n_k} \to \tilde{\theta}$. از طرفی چون $\hat{\theta}_{n_k}$، MLE است، پس

$$\frac{1}{n_k} L\left(\hat{\theta}_{n_k}\right) \geq \frac{1}{n_k} L(\theta^*),$$

با حد گیری از طرفین این نامساوی داریم

$$g\left(\tilde{\theta}\right) \geq g(\theta^*).$$

و این با فرض θ^* تنها ماکسیمم $g(\theta)$ در تناقض بـوده، پـس $\hat{\theta}_n$ به θ^* همگراست.

حالت۲: $\hat{\theta}_n$ کراندار نیست.

در این حالت نیز فرض کنید $\hat{\theta}_n$ به θ^* همگرا نباشد. بنابراین، زیـــر مجموعـــه ای ماننـــد $\{n_k\}$ از $\{n\}$ وجـــود دارد $\hat{\theta}_{n_k} \to \infty$. پس داریم

$$\frac{1}{n_k} L\left(\hat{\theta}_{n_k}\right) \geq \frac{1}{n_k} L(\theta^*),$$

در نتیجـه $\frac{1}{n_k} L\left(\hat{\theta}_{n_k}\right) \to -\infty$ (زیـرا $\frac{1}{n_k} L\left(\hat{\theta}_{n_k}\right) \to g\left(\hat{\theta}_{n_k}\right)$ و $g(\infty) = -\infty$). هم چنین $\frac{1}{n_k} L(\theta^*)$ همگرا به یک مقدار ثابت است (زیرا $\frac{1}{n_k} L(\theta^*) \to g(\theta^*)$ کـه $g(\theta^*)$ یـک مقـدار ثابـت است)، پس به این نتیجه می رسیم که $\hat{\theta}_{n_k}$، MLE نیست کـه با فرض مسئله در تناقض است. پس $\hat{\theta}_n$ به θ^* همگراست.

قضیه۳-۲: برآورد درستنمایی ماکسیمم θ یک برآورد سازگار برای θ است.

اثبات: با کمـک لـم هـای لـم ۲۰۳ و لـم ۳۰۳ و نتیجـه ۲۰۲ موجود در کتـاب بـرآورد نقطـه ای لهمـن[1] [۸]، صـفحه ٤١٤، قضیه فوق به آسانی اثبات می شود.

لم ۳-٤: فـرض کنیـد U_iهادنبالـه ای از متغیرهـای تـصادفی مستقل و هم توزیع با $E(U_1)=0$، $Var(U_1)=1$ و $N(n)$ دارای توزیع دوجمله ای (n,p) است که تابع احتمال تجمعی $N(n)$ به صورت زیر است

$$P(N(n)=i) = \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}; \quad i = 0,1,\ldots,n$$

پس برای $n \to \infty$،

$$\frac{1}{\sqrt{N(n)}} \sum_{i=1}^{N(n)} U_i \xrightarrow{d} N(0,1).$$

اثبات: فرض کنید

$$Y_{N(n)} = \frac{1}{\sqrt{N(n)}} \sum_{i=1}^{N(n)} U_i,$$

تابع مشخصه[2] $Y_{N(n)}$ را با $\phi_{N(n)}(t)$ نمایش داده که بـه صـورت زیر است،

$$\begin{aligned}\phi_{N(n)}(t) &= E\left(e^{itY_{N(n)}}\right) \\ &= \sum_{k=0}^{n} E\left(e^{it\frac{1}{\sqrt{k}}\sum_{i=1}^{k} U_i} \middle| N(n)=k\right) \\ &\quad \times \frac{n!}{i!(n-i)!} p^k (1-p)^{n-k}.\end{aligned}$$

می دانیم

$$\begin{aligned}E\left(e^{it\frac{1}{\sqrt{k}}\sum_{i=1}^{k} U_i}\right) &= E\left(e^{\sum_{i=1}^{k} it\frac{1}{\sqrt{k}}U_i}\right) \\ &= \left(\phi\left(\frac{t}{\sqrt{k}}\right)\right)^k,\end{aligned}$$

حال اگر $\phi_U(.)$ تابع مشخصه U_1 باشد، برای t ثابت داریم

$$\begin{aligned}\left|\phi_{N(n)}(t) - e^{\frac{-t^2}{2}}\right| &\leq \sum_{k=0}^{n} \left| E\left(e^{it\frac{1}{\sqrt{k}}\sum_{i=1}^{k} U_i} \middle| N(n)=k\right)\right. \\ &\quad \left. - e^{\frac{-t^2}{2}} \right| \frac{n!}{i!(n-i)!} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \left| \left(\phi\left(\frac{t}{\sqrt{k}}\right)\right)^k - e^{\frac{-t^2}{2}} \right| \\ &\quad \times \frac{n!}{i!(n-i)!} p^k (1-p)^{n-k}.\end{aligned}$$

بنابر قضیه حد مرکزی[3]

$$\frac{\overline{U} - 0}{\frac{1}{\sqrt{k}}} \xrightarrow{d} N(0,1).$$

در نتیجه

$$\frac{\sum_{i=1}^{k} U_i}{\sqrt{k}} \xrightarrow{d} N(0,1),$$

پس $\left(\phi_{U_1}(t)\right)^k \to e^{\frac{-t^2}{2}}$ یعنی

[1] Lehmann

[2] Characteristic function

[3] Central Limit Theorem

$$\lim_{k\to\infty}\left(\phi_U\left(\frac{t}{\sqrt{k}}\right)\right)^k = e^{\frac{-t^2}{2}}.$$

بنابراین برای $\varepsilon>0$ داده شده، $N_1(t)$ را به اندازه کافی بـزرگ اختیار می کنیم پس برای $k \geq N_1(t)$

$$\left|\left(\phi_U\left(\frac{t}{\sqrt{k}}\right)\right)^k - e^{\frac{-t^2}{2}}\right| \leq \varepsilon.$$

علاوه بر این، برای $N_1(t)$ ثابت، n را به اندازه کـافی بـزرگ اختیار می کنیم پس طبق قضیه حد مرکزی

$$\sum_{i=0}^{N_1(t)} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \leq \varepsilon.$$

بنابراین

$$\left|\phi_{N(n)}(t) - e^{\frac{-t^2}{2}}\right| \leq 3\varepsilon.$$

حـال چـون ε دلخـواه اسـت و $e^{\frac{-t^2}{2}}$ تـابع مشخـصه توزیـع $N(0,1)$ است پس

$$\frac{1}{\sqrt{N(n)}} \sum_{i=1}^{N(n)} U_i \xrightarrow{d} N(0,1).$$

پس لم فوق ثابت می شود.

قضیه۳-۳: برآورد درستنمایی ماکسیمم دارای توزیـع مجـانبی زیر است

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{c^2}\right),$$

که

$$\sigma^2 = \left[E\left\{\left(T - \frac{1}{\theta_0}\right)^2 | T \notin (L,R)\right\} - \left(E\left\{\left(T - \frac{1}{\theta_0}\right) | T \notin (L,R)\right\}^2 \right)\right] + \left[E\left\{L^2 | T \in (L,R)\right\} - \left(E\left\{L | T \in (L,R)\right\}\right)^2\right] + \left[E\left\{V^2 | T \in (L,R)\right\} - \left(E\left\{V | T \in (L,R)\right\}\right)^2\right]$$

و در آن

$$V = \frac{Ze^{-\theta_0 Z}}{1-e^{-\theta_0 Z}},$$

و

$$c = \frac{p(\theta_0)}{\theta_0^2} + (1 - p(\theta_0)) \times \left\{ E\left\{ \frac{Z^2 e^{-\theta_0 Z} Z}{(1-e^{-\theta_0 Z})^2} | T \in (L,R) \right\}\right\}.$$

اثبات: می دانیم

$$L(\theta) = n_1 \ln\theta - \theta \sum_{i=1}^{n_1} T_i - \theta \sum_{i=n_1+1}^{n_1+n_2} L_i + \sum_{i=n_1+1}^{n_1+n_2} \ln\left(1 - e^{-\theta Z_i}\right),$$

پس

$$L'(\theta) = -\frac{n_1}{\theta} - \sum_{i=1}^{n_1} T_i - \sum_{i=n_1+1}^{n_1+n_2} L_i + \sum_{i=n_1+1}^{n_1+n_2} \frac{Z_i e^{-\theta Z_i}}{\left(1 - e^{-\theta Z_i}\right)},$$

و

$$L''(\theta) = \frac{n_1}{\theta^2} + \sum_{i=n_1+1}^{n_1+n_2} \frac{Z_i^2 e^{-\theta Z_i}}{\left(1 - e^{-\theta Z_i}\right)^2}.$$

با استفاده از قضیه مقدار میانگین،

$$L'(\hat{\theta}) - L'(\theta_0) = (\hat{\theta} - \theta_0) L''(\bar{\theta}),$$

که در آن $\bar{\theta}$ یک نقطه بین $\hat{\theta}$ و θ_0 است. بنابراین،

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\frac{1}{\sqrt{n}} L'(\theta_0)}{\frac{1}{n} L''(\bar{\theta})}.$$

اثبات کامل می شود اگر نشان دهیم

$$\frac{1}{\sqrt{n}} L'(\theta_0) \xrightarrow{d} N(0,1), \qquad (18)$$

و

$$\frac{1}{n} L''(\bar{\theta}) \longrightarrow c. \quad a.s. \qquad (19)$$

بنابر لم ٤ رابطه (۱۸) برقرار است. از طرفی بنابر قـانون قـوی اعداد بزرگ $\bar{\theta} \to \theta_0$ $a.s.$، در نتیجه فرمول (۱۹) نیـز ثابـت می شود.

۴- نتیجه

در این مقاله ابتدا سانسور میانی را بـه طـور دقیـق تعریـف کردیم و سپس به بررسی استنباطی توزیع نمایی وقتی داده هـا از سانسور میـانی آمـده انـد پـرداختیم. بـرای محاسـبه MLE پارامتر θ از روش عددی و الگوریتم EM اسـتفاده کـردیم و سپس نشان دادیم روش های تکراری که در فرمول (۷) مطرح شده است همگراست اگر رابطه (۱۲) برقـرار باشـد. در ادامـه ثابت کردیم $\frac{1}{n} L(\theta) \to g(\theta)$ $a.s.$ و $g(\theta)$ یک تابع تک نمایی با ماکسیمم واحد است. در بخـش ۳، بـا بیـان لـم ۳-۳ ثابـت کردیم که برآورد درستنمایی ماکسیمم θ یک برآورد سـازگار

برای θ_0 است و نشان دادیم که $\sqrt{n}(\hat{\theta}-\theta_0)\xrightarrow{d} N\left(0,\frac{\sigma^2}{c^2}\right)$ می باشد.

منابع

[1]. Babu, G.J., Rao, C.R. and Rao, M.B. (1992). Nonparametric estimation of specific exposure rate in risk and survival analysis. Journal of American Statistical Association, 87, 84-89.

[2]. Chow, Y.S. and Teicher, H. (1980). Probability Theory, Springer, New York.

[3]. Iyer, S.K., Kundu, D. and Jammala-madaka, S.R. (2008). Analysis of Middle Censored Data with Exponential Lifetime Distribution. Journal of Statistical Planing and Inference, to be appear.

[4]. Jammalamadaka, S.R. and Mangalam, V. (2003). Nonparametric estimation for middle censored data. Journal of Naonparametric Statistics, 15, 253-265.

[5]. Jammalamadaka, S.R and Iyer, S.K. (2004). Approximate self consistency for middle censored data, Journal of Statistical Planing and Inference, 124, 75-86.

[6]. Kaplan, F.L. and Meier, P. (1958). Nonparametric estimation from incomplete observation, Journal of American Statistical Association, 63, 457-481.

[7]. Thompson; J., (1999) Simulation: A Modeler's Approach. Wiley, New York, 306 pages.

[8]. Lehmann, E.L. (1986). Theory of point estimation. Wiley.

[۹]. اسماعیلی، ح. (۱۳۸٤)، آنالیز عـددی مبتنـی بـر مطلـب، انتشارات بو علی سینا، همدان.

[۱۰]. عالم زاده، ع. (۱۳۷٤)، حساب دیفرانسیل و انتگـرال با هندسه تحلیلی، موسسه نشر علوم نوین.