آشنایی با روش دلتا برای تعیین توزیع‌های مجانبی و کاربردهایی از آن

منوچهر خردمندنیا[۱]

چکیده

فرض کنید که یک آماره دارای توزیع مجانبی نرمال است. آگرستی (۱۹۹۰) نشان داد که توابع بسیاری از آن آماره نیز به طور مجانبی نرمال هستند. وی یک روش ساده برای به دست آوردن این توزیع‌ها ارایه نمود. در مقاله حاضر کاربردهایی را از روش مذکور، در الگوسازی لگاریتم خطی و در الگوسازی لجیت معرفی می‌کنیم.

۱ مقدمه

برای تعیین توزیع مجانبی تابعی از یک آماره می‌توان تحت شرایط خاصی از روش دلتا استفاده کرد. در این مقاله ضمن معرفی این روش، شرایط خاص مذکور را روشن می‌کنیم. این روش کاربردهای فراوان دارد. در مقاله حاضر بعضی از کاربردهای این روش را که در تحلیل‌های چندمتغیره گسسته به‌کار می‌آیند معرفی می‌کنیم. از مراجع مهم مربوط به این روش می‌توان از رائو (۱۹۷۳) و آگرستی (۱۹۹۰) نام برد. بوت استرپ روش دیگری برای برآورد واریانس و توزیع نمونه‌ای آماره‌هاست. برای آشنایی با این روش به ایران‌پناه و پاشا(۱۳۷۶) مراجعه کنید. جک نایف نیز روش دیگری برای برآورد یک برآوردگر است. برای آشنایی با این روش به نیرومند(۱۳۷۶) مراجعه کنید. روشهای بوت‌استرپ و جک نایف، به‌خصوص در مواردی که شرایط لازم برای به‌کارگیری روش دلتا موجود نیست مفیدند.

[۱] دکتر منوچهر خردمندنیا، گروه آمار، دانشگاه اصفهان

۲ روش دلتا برای تعین توزیع مجانبی

یک دنباله X_n از متغیرهای تصادفی را به طور مجانبی نرمال (AN) با میانگین μ_n و انحراف معیار $\sigma_n > ۰$ گویند، هرگاه $(X_n - \mu_n)/\sigma_n \xrightarrow{d} N(۰, ۱)$، یعنی هرگاه تابع چگالی احتمال $(X_n - \mu_n)/\sigma_n$ به توزیع $N(۰, ۱)$ همگرا باشد. در این صورت می‌نویسیم $X_n \sim AN(\mu_n, \sigma_n^۲)$.

فرض کنید T_n یک آماره است که بستگی به n (اندازه نمونه) دارد و برای nهای بزرگ تقریباً دارای توزیع نرمال با میانگین θ و انحراف معیار $\sigma/\sqrt{n}$ است. به عبارت دقیق‌تر، وقتی $n \to \infty$، تابع چگالی احتمال $\sqrt{n}(T_n - \theta)/\sigma$ به توزیع $N(۰, ۱)$ همگراست. یعنی

$$T_n \sim AN(\theta, \sigma^۲/n) \cdot \qquad (۱)$$

تحت شرایط فوق اگر $g(T_n)$ تابعی از T_n باشد به‌طوری که $g(\theta)$ حداقل دوبار برحسب θ مشتق‌پذیر باشد آنگاه

$$g(T_n) \sim AN[g(\theta), \frac{\sigma^۲}{n}(g'(\theta))^۲] \qquad (۲)$$

که در آن $g'(\theta) = dg(\theta)/d\theta$. برای اثبات نتیجه (۲) ملاحظه کنید که بسط تیلور تابع $g(t)$ حول $t = \theta$ عبارت است از

$$g(t) = g(\theta) + (t-\theta)g'(\theta) + (t-\theta)^۲ g''(\theta^*)/۲$$

که در آن $t \leq \theta^* \leq \theta$ و $g''(\theta)$ مشتق دوم $g(\theta)$ در نقطه $\theta = \theta^*$ است. در صورتی که در رابطه اخیر متغیر تصادفی T_n را به جای t قرار دهیم، داریم:

$$\begin{aligned}\sqrt{n}[g(T_n) - g(\theta)] &= \sqrt{n}(T_n - \theta)g'(\theta) \\ &+ \sqrt{n}(T_n - \theta)^۲ g''(\theta^*)/۲ \cdot\end{aligned}$$

اما وقتی $n \to \infty$، جمله دوم یعنی $\sqrt{n}(T_n - \theta)^۲ g''(\theta^*)/۲$ در احتمال به صفر می‌گراید. بنابراین دو متغیر تصادفی $\sqrt{n}[g(T_n) - g(\theta)]$ و $\sqrt{n}(T_n - \theta)g'(\theta)$ دارای توزیع حدی یکسانی هستند و لذا نتیجه (۲) حاصل می‌شود.

با توجه به اینکه $\sigma^۲ = \sigma^۲(\theta)$ و $g'(\theta)$ به‌طور کلی به پارامتر نامعلوم θ بستگی دارند، واریانس مجانبی $g(T_n)$ نیز نامعلوم است. وقتی که $g'(\cdot)$ و $\sigma = \sigma(\cdot)$ در نقطه θ پیوسته باشند، $\sigma(T_n)g'(T_n)$ یک برآوردگر سازگار برای $\sigma(\theta)g'(\theta)$ است، بنابراین در این حالت

$$g(T_n) \sim AN[g(\theta), \frac{\sigma^۲(t_n)}{n}(g'(t_n))^۲] \qquad (۳)$$

لذا فاصله زیر، یک فاصله اطمینان ۹۵ درصد برای $g(\theta)$ است

$$g(t_n) \pm ۱/۹۶\ \sigma(t_n) \mid g'(t_n) \mid /\sqrt{n} \qquad (۴)$$

که در آن t_n یافته‌ای از T_n است. برای توضیحات نظری بیشتر به فصل ۱۲ کتاب آگرستی (۱۹۹۰) مراجعه کنید. در این مرجع، نتیجه (۲)، روش دلتا برای تعیین توزیع مجانبی نامیده شده است. در همین مرجع، تاریخچه‌ای طولانی از فرایند تکامل این روش ارایه شده است.

برای یافتن وجه تسمیه اصطلاح روش دلتا، با بررسی متون آماری در دسترس، به دایرة‌المعارف اصطلاحات آماری می‌رسیم: در این مرجع، اصطلاح روش دلتا مستقیماً تعریف و تشریح نشده است. ولی در صفحه ۶۴۶ از جلد هشتم این دایرة‌المعارف، با ارجاع به فصل ۱۴ کتاب بیشاپ و همکاران (۱۹۷۵) از روش دلتا به عنوان نامی جانشین برای روش دیفرانسیل آماری نام برده است. در این کتاب، روش دیفرانسیل آماری به عنوان روشی برای دستیابی به تقریبی برای مقدار مورد انتظار تابعی‌از یک متغیر تصادفی تعریف شده و آمده است که انحراف متغیرهای تصادفی را از مقادیر مورد انتظارشان $(X_j - \mu_j)$، که در روش دیفرانسیل آماری به‌کار می‌آید، اغلب با نماد δX_j نشان می‌دهند و لذا وجه تسمیه روش دلتا حاصل می‌شود. روشی که آگرستی (۱۹۹۰) تحت عنوان روش دلتا برای تعیین توزیع مجانبی معرفی نموده، شکل تعدیل و تکامل یافته‌ای از روش دیفرانسیل آماری (روش دلتا) است. در بخش بعدی، به منظور رعایت اختصار به‌جای اصطلاح دقیق ولی طولانی روش دلتا برای تعیین توزیع مجانبی، اصطلاح روش دلتا را به‌کار می‌بریم.

۳ کاربردهایی از روش دلتا برای تعیین توزیع مجانبی

(i) فرض کنید متغیر تصادفی Y مجموع n پواسون مستقل هر یک با میانگین $\mu > 0$ است. بنابراین $Y \sim Poisson(n\mu)$ و می‌خواهیم توزیع مجانبی هر یک از آماره‌های زیر را به دست آوریم.

(الف) $\ln(Y)$ (ب) $1/Y$ (ج) $\sqrt{Y}$

براساس قضیه حد مرکزی می‌دانیم که $Y/n \sim AN[\mu, \mu/n]$.

(الف) برای تعیین توزیع مجانبی $\ln(Y)$ ابتدا توزیع مجانبی $\ln(Y/n)$ را تعیین می‌کنیم. در اینجا بر اساس نمادهای اختصاری بخش قبل داریم

$$T_n = Y/n \quad , \quad \theta = \mu$$

$$\sigma^2(\theta) = \sigma^2(\mu) = \mu$$

$$g(\theta) = \ln(\theta) = \ln(\mu)$$

$$\frac{\sigma^2(\mu)}{n}(g'(\mu))^2 = \frac{\mu}{n}(\frac{1}{\mu})^2 = \frac{1}{n\mu} .$$

بنابراین بر اساس روش دلتا می‌توان نوشت

$$\ln(Y/n) \sim AN[\ln(\mu), 1/n\mu] .$$

از رابطه‌ی اخیر بلافاصله رابطه زیر نتیجه می‌شود:

$$\ln(Y) \sim AN[\ln(m), 1/m] \qquad (5)$$

که در آن $m = E(Y) = n\mu$. بنابراین فاصله $\ln(y) \pm 1.96(1/y)$ یک فاصله اطمینان ۹۵ درصد برای $\ln(m)$ است که در آن y یافته‌ای از Y و برآورد ML پارامتر m می‌باشد.
با استفاده از (۵) می‌توان نسبت به پارامترهای الگوی خطی اشباع‌شده، بدون توسل به روشهای عددی، استنباط آماری انجام داد. از آنجا که تکیه‌گاه Y شامل صفر است، واریانس دقیق $\ln(Y)$ وجود ندارد. ولی با افزایش n احتمال $Y = 0$ به‌سرعت به صفر می‌گراید.

(ب) برای تعیین توزیع مجانبی $1/Y$ ابتدا توزیع مجانبی n/Y را تعیین می‌کنیم. در اینجا داریم $T_n = Y/n$ و $g(\theta) = 1/\theta = 1/\mu$. لذا بر اساس روش دلتا می‌توان نوشت

$$n/Y \sim AN[1/\mu, 1/n\mu^3] .$$

از رابطه‌ی اخیر به‌سهولت رابطه‌ی زیر نتیجه می‌شود:

$$1/Y \sim AN(1/m, 1/m^3) . \qquad (6)$$

بنابراین فاصله $(1/y) \pm 1.96(1/y)^{1.5}$ یک فاصله اطمینان ۹۵ درصد برای $1/m$ می‌باشد. که در آن y یافته‌ای از Y و برآورد ML پارامتر $m = n\mu$ است.

اگر برای داده‌های یک جدول توافقی، یک الگوی اشباع شده‌ی خطی را با پاسخ پوآسون و تابع پیوند وارون در نظر گیریم، با استفاده از (۶)، بدون توسل به روشهای عددی، می‌توان نسبت به پارامترهای الگو استنباط آماری انجام داد.

(ج) برای تعیین توزیع مجانبی $\sqrt{Y}$ ابتدا توزیع مجانبی $\sqrt{Y/n}$ را تعیین می‌کنیم. در اینجا داریم $T_n = Y/n$ و $g(\theta) = \sqrt{\theta} = \sqrt{\mu}$. لذا بر اساس روش دلتا می‌توان نوشت

$$\sqrt{Y/n} \sim AN[\sqrt{\mu}, 1/4n]$$

در نتیجه

$$\sqrt{Y} \sim AN[\sqrt{m}, 1/4] \qquad (7)$$

ملاحظه می‌شود که واریانس مجانبی $\sqrt{Y}$ عددی ثابت است. بنابراین برای پاسخ‌های پواسون، تبدیل ریشه دوم طریقه‌ای برای تثبیت واریانس است.

(ii) فرض کنید متغیر تصادفی Y مجموع n برنولی مستقل هر یک با احتمال موفقیت $\pi > 0$ است. بنابراین $Y \sim Bin(n, \pi)$ و می‌خواهیم توزیع مجانبی هر یک از دو آماره زیر را به‌دست آوریم.

(الف) $\ln \frac{Y}{n-Y}$ (ب) $\sin^{-1}(\sqrt{Y/n})$

براساس قضیه حد مرکزی می‌دانیم که برای $0 < \pi < 1$ داریم

$$\frac{Y}{n} \sim AN[\pi, \frac{\pi(1-\pi)}{n}] \cdot$$

(الف) برای تعیین توزیع مجانبی $\ln(Y/(n-Y))$ بر اساس نمادهای اختصاری بخش ۲ داریم

$$Tn = Y/n \quad , \quad \theta = \pi$$

$$\sigma^2(\theta) = \sigma^2(\pi) = \pi(1-\pi)$$

$$g(\theta) = g(\pi) = \pi/(1-\pi) \cdot$$

بنابراین بر اساس روش دلتا داریم

$$\ln(\frac{Y}{n-Y}) \sim AN[\ln(\frac{\pi}{1-\pi}), \frac{1}{n\pi(1-\pi)}] \quad (8)$$

لذا فاصله زیر یک فاصله اطمینان ۹۵ درصد برای $Logit(\pi) = \ln(\pi/(1-\pi))$ است

$$\ln(\frac{y}{n-y}) \pm 1/96 \frac{n}{y(n-y)}$$

که در آن y یافته‌ای از Y و برآورد ML پارامتر $n\pi$ است.

بااستفاده از (۸)، بدون توسل به روشهای عددی می‌توان راجع به پارامترهای الگوی اشباع‌شده‌ی لجیت با متغیرهای توضیحی رسته‌ای استنباط آماری انجام داد. از آنجا که احتمال $Y = 0$ یا $Y = n$ مثبت است، لجیت می‌تواند برابر $-\infty$ یا $+\infty$ شود. ولی با افزایش n، احتمال $Y = 0$ یا $Y = n$ به‌سرعت به صفر می‌گراید.

(ب) برای تعیین توزیع مجانبی $\sin^{-1}(\sqrt{Y/n})$ داریم $g(\theta) = \sin^{-1}(\sqrt{\theta}) = \sin^{-1}(\sqrt{\pi})$.

بنابراین بر اساس روش دلتا می‌توان نوشت

$$\sin^{-1}(\sqrt{Y/n}) \sim AN[\sin^{-1}(\sqrt{\pi}), 1/4n] \cdot \quad (9)$$

ملاحظه می‌شود که برای پاسخ‌های دوجمله‌ای تبدیل $\sin^{-1}(\sqrt{Y/n})$ طریقه‌ای برای تثبیت واریانس است.

مراجع

[1] Agresti, A., *Categorical Data Analysis*, John Wiley & Sons, (1990).

[2] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W., *Discrete Multivariate Analysis*, MIT Press, Cambridge, MA, (1975).

[3] *Encyclopedia of Statistical Sciences*, Vol. 8, John Wiley & Sons, (1981).

[4] Rao, C. R., *Linear Statistical Interface and its Applications*, 2nd Ed., John Wiley & Sons, (1973).

[۵] ایران‌پناه، ن. و پاشا، ع. (۱۳۷۶)، آشنایی با الگوریتم بوت استرپ، اندیشه آماری، سال دوم، شماره اول.

[۶] نیرومند، ح. (۱۳۷۶)، آشنایی با جک نایف، اندیشه آماری، سال دوم، شماره دوم.

یک نشریه علمی انجمن پزشکان قلب آمریکا[۱]، آماری را پیرامون استفاده غیرصحیح از علم آمار ارائه داده‌است که جای تأمل دارد ! در این نشریه آمده است که:

«تقریباً نیمی از مقالاتی که در نشریات علمیِ پزشکی چاپ شده و در آنها تجزیه و تحلیل آماری انجام گرفته است، متاسفانه از روشهای نادرست استفاده نموده‌اند.

این مقاله باعث شد که هیأت تحریریه نشریه در داوری مقالات بیشتر دقت کنند، امّا این سؤال باقی مانده است که آیا سایر نشریات هم چنین خواهند کرد؟

بسیار جای تأسف است که محققین علم آمار با استفاده از پیشرفته‌ترین مباحث ریاضی از جمله؛ نظریه اندازه، توپولوژی، آنالیز تابعی و فرایندهای تصادفی درصدد پیشبرد علم آمار و تقویت مبانی این علم هستند و از طرف دیگر مقالات علمی هنوز در استفاده از روشهای ساده‌ای مانند آزمون t دچار اشتباه می‌شوند. تأسف‌بارتر این است که نه تنها دانشمندان سایر علوم آزمونهای آماری را صحیح به کار نمی‌برند، بلکه حتی آمارشناسان نیز چنین خطاهایی را مرتکب می‌شوند!»

[1] Glantz (1980), *How to Detect, Correct and Prevent Errors in the Medical Literature*, Biostatistics , Vol. 61, 1-7.