

کارآیی بین میانگین و میانه در نمونه‌های بزرگ از چند توزیع پیوسته مهم

علیرضا سلیمانی خیری^۱

چکیده

هدف از این نوشتار، پرداختن به کارآیی مجانبی میانگین حسابی و میانه دو معیار مرکزی نمونه‌ای، با توجه به ساختار تابع چگالی مورد نظر و بررسی برتری هر کدام نسبت به دیگری و سرانجام استباط یک نتیجه کلی می‌باشد.

۱ مقدمه

نخست به بیان چند تعریف می‌پردازیم:

تعریف ۱-۱: \bar{X}_n را میانگین حسابی نمونه X_1, \dots, X_n می‌خوانیم هرگاه:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

تعریف ۱-۲: نماد \bar{X}_n را میانه نمونه‌ای بر اساس نمونه

X_1, \dots, X_n ، و $\hat{\mu}$ را میانه جامعه می‌نامیم.

تعریف ۱-۳: ϕ_n را یک برآورده ناریب از θ می‌خوانیم چنانچه برای تمام مقادیر θ :

$$E_\theta(\phi_n) = \theta$$

تعریف ۱-۴: ϕ_n برآورده ناریب سازگار (به طور ضعیف) برای

^۱ علیرضا سلیمانی خیری، گروه ریاضی، دانشگاه تربیت معلم سبزوار

^۲ $p \in (0, 1)$

^۳ از تعریف ۱-۷ نتیجه می‌شود که X حول α نسبت به Y متتمرکزتر است هرگاه $var_\alpha(X) \leq var_\alpha(Y)$.

حال توزیع $Y = F(\hat{\xi}_p)$ را می‌پاییم که به قرار زیر خواهد بود:

$$g_Y(y) = \frac{n!}{(m-1)!(n-m)!} y^{m-1} (1-y)^{n-m}; 0 < y < 1$$

با تعریف $Z = \frac{\sqrt{n}(Y-p)}{\sqrt{p}(1-p)}$ ، خواهیم داشت:^۵

$$\begin{aligned} f_Z(z) &= \frac{1}{\sqrt{n}} \cdot \frac{n!}{(m-1)!(n-m)!} \\ &\times p^{m-\frac{1}{2}} q^{n+m-\frac{1}{2}} \\ &\times [1+z\sqrt{\frac{q}{np}}]^{m-1} [1-z\sqrt{\frac{p}{nq}}]^{n-m} \end{aligned}$$

با میل دادن $\rightarrow \infty$ ، خواهیم داشت: $z < -\infty$. با توجه به این و ساختار متغیر تصادفی Z ، نشان می‌دهیم که دارای توزیع حدی نرمال است. برای سادگی از طرفین $f_Z(z)$ نسبت به z لگاریتم می‌گیریم.

$$\begin{aligned} \log f_Z(z) &= \log \left(\frac{1}{\sqrt{n}} \cdot \frac{n!}{(m-1)!(n-m)!} p^{m-\frac{1}{2}} q^{n+m-\frac{1}{2}} \right) \\ &+ (m-1) \log [1+z\sqrt{\frac{q}{np}}] \\ &+ (n-m) \log [1-z\sqrt{\frac{p}{nq}}] \end{aligned}$$

با میل دادن $\rightarrow \infty$ ، جمله اول که مستقل از Z است به مقدار ثابت مثلاً C میل می‌نماید. اما با استفاده از بسط مکلورن $\log(1+x) \approx x$ و $\log(1-x) \approx -x$ و توجه به اینکه $m = np$ داریم:

$$(m-1) \log [1+z\sqrt{\frac{q}{np}}] + (n-m) \log [1-z\sqrt{\frac{p}{nq}}] \xrightarrow[n \rightarrow \infty]{} -\frac{z^2}{2}$$

$$\log f_Z(z) \xrightarrow[n \rightarrow \infty]{} C(-\frac{z^2}{2})$$

بنابراین:

در نتیجه:

$$f_Z(z) \xrightarrow[n \rightarrow \infty]{} de^{-\frac{z^2}{2}}; \quad \text{ثابت } d$$

متغیر تصادفی X حول مثلاً α نسبت به Y متتمرکزتر (more concentrated) یا به طور اختصار m.co می‌باشد، هرگاه برای $t > 0$

$$Pr(|X - \alpha| \leq t) \geq Pr(|Y - \alpha| \leq t)$$

تعريف ۱-۸: اگر برای برآورده ϕ_n از θ ، $\sqrt{n}(\phi_n - \theta)$ به طور مجانبی دارای توزیع نرمال با میانگین صفر و واریانس متناهی (تابعی نامنفی از θ) باشد، آنگاه ϕ_n یک برآورده سازگار به طور مجانبی نرمال (به طور اختصار CAN) برای θ نامیده می‌شود.

۲ روابط

حال به بیان دو قضیه که در ادامه بحث مورد استفاده قرار می‌گیرند، می‌پردازیم.

قضیه ۱-۲: فرض کنیم X_1, X_2, \dots, X_n متغیر تصادفی مستقل و همتوزیع با تابع توزیع $F_X(\cdot)$ و چگالی پیوسته $f_X(\cdot)$ باشند که در \mathbb{R}^p تعریف شده‌اند، آنگاه:

$$\sqrt{n}(\hat{\xi}_p - \xi_p) \xrightarrow{L} N(0, \frac{p(1-p)}{[f(\xi_p)]^2})$$

اثبات: داریم

$$\begin{aligned} Pr(\hat{\xi}_p \leq x) &= Pr(\text{مشاهدہ } np \text{ حداقل مشاهده } x) \\ &= \sum_{r=np}^n \frac{n!}{r!(n-r)!} [F(x)]^r [1-F(x)]^{n-r} \end{aligned}$$

حال اگر np صحیح باشد $m = np$ در غیر اینصورت $m = [np] + 1$. بنابراین^۴

$$Pr(\hat{\xi}_p \leq x) = \frac{n!}{(m-1)!(n-m)!} \int_0^{F(x)} t^{m-1} (1-t)^{n-m} dt$$

در نتیجه با مشتق‌گیری نسبت به x داریم:

$$g_{\hat{\xi}_p}(x) = \frac{n!}{(m-1)!(n-m)!} [F(x)]^{m-1} [1-F(x)]^{n-m} f(x)$$

^۴ با انتگرال گیری جزء به جزء مکرر از انتگرال بنای ناقص، تساوی محرز می‌گردد.
^۵ $q = 1 - p$

توجه به قضیه (۲-۲) و با توجه به تقارن حول $\bar{\mu} = \mu$ و \bar{X}_n

هر دو برآوردهای ناریب و سازگار باند. بعلاوه:

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$$

و با توجه به قضیه (۱-۲):

$$\bar{X}_n \xrightarrow{L} N(\mu, \frac{\pi\sigma^2}{2n})$$

که برای n های بزرگ مشاهده می شود، \bar{X}_n همچنان m.co. حول μ نسبت به \bar{X}_n است. همچنین:

$$e(\bar{X}_n, \bar{X}_n) = \frac{var(\bar{X}_n)}{var(\bar{X}_n)} \simeq \frac{2}{\pi} < 1$$

از آنجا که $\frac{2}{\pi} = 0.64$. پس می توان گفت \bar{X}_n در نمونه های بزرگ μ را با ۶۴٪ مشاهدات برآورد می نماید. چه موقع μ را توسط \bar{X}_n برآورد می نماییم؟

با توجه به اینکه میانگین حسابی تحت تأثیر مشاهدات بوده، ولی میانه حساسیت کمتری به مشاهدات در قیاس با میانگین داشته و تنها به یک یا دو داده وسطی (بعد از مرتب شدن به ترتیب افزایش داده ها) سروکار دارد، پس برای نمونه های بزرگ همانطور که مشاهده نمودیم \bar{X}_n قابلیت اعتماد بیشتری نسبت به \bar{X}_n جهت برآورد μ دارد.

گاهی اوقات، استفاده از \bar{X}_n می تواند به این علت باشد که بخواهیم ترجیحاً $EX = 0.5$ و نه EX را برآورد نماییم، زیرا در حالت کلی $EX \neq 0.5$. بنابراین \bar{X}_n یک برآورده مناسب خواهد بود. تحت شرایط کلی، \bar{X}_n یک برآورده سازگار و CAN از 0.5 است.

۲-۳ توزیع کشی

فرض نماییم X_1, X_2, \dots, X_n یک نمونه تصادفی از چگالی زیر باشند.

$$f_X(x; \theta) = \frac{1}{\pi[1 + (x - \theta)^2]}, \quad x \in R, \theta > 0$$

^۶ برای جزئیات بیشتر در مورد اثبات به صفحات ۳۸۵-۳۸۶ منبع [۴] مراجعه نمایید.

یعنی توزیع حدی $Z = (\bar{X}_n - \mu)/\sqrt{n}$ است.

پس $p = F(\xi_p)$ و $Y = F(\hat{\xi}_p)$ و $\sqrt{n}(Y - p) \xrightarrow{L} N(0, pq)$. اما

$$\sqrt{n}(F(\hat{\xi}_p) - F(\xi_p)) \xrightarrow{L} N(0, pq)$$

بنابراین: از طرفی با توجه به بسط تیلور $F(Y) \approx F(p) + f(p)(Y - p)$ داریم:

$$F(\hat{\xi}_p) - F(\xi_p) = f(\xi_p)(\hat{\xi}_p - \xi_p) + \epsilon_n$$

با توجه به اینکه وقتی $\xi_p \xrightarrow{P} \hat{\xi}_p$ ، $\epsilon_n \xrightarrow{P} 0$. پس:

$$\sqrt{n}f(\xi_p)(\hat{\xi}_p - \xi_p) \xrightarrow{L} N(0, pq)$$

و سرانجام:

$$\sqrt{n}(\hat{\xi}_p - \xi_p) \xrightarrow{L} N(0, \frac{pq}{[f(\xi_p)]^2})$$

قضیه ۲-۲: اگر X_1, X_2, \dots, X_n متغیر تصادفی مستقل و هم توزیع با تابع توزیع $F_X(x; \theta)$ باشند آنگاه با فرض موجود بودن $E_\theta(X)$.

(۱) \bar{X}_n برآورده ناریب و سازگار برای $E_\theta(X)$ است.

(۲) \bar{X}_n برآورده سازگار برای μ بوده که در حالت کلی ناریب نمی باشد، مگر توزیع حول μ متقاضن باشد.

برای اثبات می توانید به منبع شماره [۱] مراجعه نمایید.

۳ کارایی

حال به بررسی کارایی مجانية بین \bar{X}_n و \bar{X}_n در توزیعهای پیوسته خاص می پردازیم. به عنوان محور اصلی بحث، توزیع نرمال را نخست بررسی نموده و از طریق گرایش برخی توزیعها، حتی نامتقارن پیوسته را به توزیع نرمال تحت شرایط خاص نشان می دهیم.

۱-۳ توزیع نرمال

اگر X_1, X_2, \dots, X_n نمونه ای تصادفی از $N(\mu, \sigma^2)$

علوم) باشند، می دانیم که $\bar{\mu} = \mu$ (بدلیل متقاضن بودن). با

^۶ برای جزئیات بیشتر در مورد اثبات به صفحات ۳۸۵-۳۸۶ منبع [۴] مراجعه نمایید.

بنابراین:

$$e(\bar{X}_n, \bar{X}_n) \approx \frac{\pi^2}{12} = 0.82$$

که بیانگر نالاریبی و سازگاری و m.co. \bar{X}_n برای θ می‌باشد.

۵-۲ توزیع نمایی مکرر(لابلس)

فرض کنیم X_1, \dots, X_n نمونه‌ای تصادفی از توزیع زیر باشند (β معلوم).

$$f_X(x; \theta) = \frac{1}{2\beta} \exp\left\{-\frac{|x - \theta|}{\beta}\right\}; x \in R, \theta \in R, \beta > 0$$

داریم:

$$\mu = E\bar{X}_n = \theta, \quad var(\bar{X}_n) = \frac{2\beta^2}{n}$$

$$\xi_{0.5} = \tilde{\mu} = \theta, \quad var(\tilde{X}_n) \approx \frac{\beta^2}{n}$$

بنابراین:

$$e(\bar{X}_n, \tilde{X}_n) \approx 2 > 1$$

یعنی \bar{X}_n برآوردهای نالاریب، سازگار و m.co. برای θ می‌باشد و بیانگر دنباله طولانی (دم طولانی) در توزیع مورد نظر می‌باشد.

بنابراین، در توزیعهای متقارن تک نمایی، عموماً می‌توان چنین استنباط نمود که هر چه دنباله توزیع طولانی تر باشد کارآیی میانه بالاتر از میانگین و هرچه کمتر باشد، کارآیی میانگین بالاتر از میانه خواهد شد. همانگونه که مشاهده شد در توزیعهای نرمال، کشی، استونت t ، لجستیک و نمایی مکرر این موضوع کاملاً مشهود بود.

با توجه به قضیه (۱-۱) داریم:

$$var(\hat{\xi}_{0.5}) = var(\tilde{X}_n) \approx \frac{1}{4[f(\tilde{\mu})]^2}$$

از طرفی می‌دانیم که

$$var(\bar{X}_n) = \frac{\sigma^2}{n}$$

در این حالت EX وجود ندارد، پس \bar{X}_n نمی‌تواند برآوردهای مناسب برای θ باشد. از این رو \tilde{X}_n جانشینی مناسب برای θ خواهد بود، زیرا $\theta = 0.5$ ، یعنی θ میانه و \bar{X}_n برای θ می‌باشد. در مقایسه توزیع کشی با توزیع نرمال، هر چند هر دو دارای توزیع زنگولهای شکل هستند، اما نرخ همگرایی به صفر در دو چگالی متفاوت است. به عبارت دیگر رفتار دمهای دو توزیع متفاوت است. از طرفی، چون در توزیع کشیتابع مشخصه میانگین نمونه $\exp[-|t| + it\theta]$ برابر تابع مشخصه خود جامعه است، یعنی تابع میانگین نمونه با توزیع جامعه یکسان است، لذا مهم نیست که نمونه چقدر بزرگ باشد و این خود گواه بر عدم سازگاری \bar{X}_n برای θ است.

۳-۳ توزیع استودنت t

در این توزیع اگر $\nu \rightarrow +\infty$ ، آنگاه توزیع استودنت t به توزیع نرمال میل می‌نماید. بر اساس یک نمونه تصادفی n تایی از این توزیع داریم:

$$var(\bar{X}_n) = \frac{\nu}{n(\nu - 2)}; \quad \nu > 2$$

$$var(\tilde{X}_n) = \frac{\pi\nu[\Gamma(\frac{\nu}{2})]^2}{4n[\Gamma(\frac{\nu+1}{2})]^2}$$

در نتیجه:

ν	∞	10	8	5	4	3	
	$0/64$	$0/76$	$0/8$	$0/96$	$1/12$	$1/62$	$e(\bar{X}_n, \tilde{X}_n)$

۴-۳ توزیع لجستیک (Logistic Distribution)

فرض کنیم X_1, \dots, X_n یک نمونه تصادفی از توزیع زیر باشند (β معلوم).

$$f_X(x; \theta) = \frac{\exp(\frac{x-\theta}{\beta})}{\beta[1 + \exp(\frac{x-\theta}{\beta})]^2}; x \in R, \beta > 0, \theta \in R$$

داریم:

$$\mu = \tilde{\mu} = \xi_{0.5} = \theta$$

$$var(\bar{X}_n) = \frac{\beta^2 \pi^2}{2n}, \quad var(\tilde{X}_n) \approx \frac{4\beta^2}{n}$$

^۷ می‌دانیم که MLE نیز میانه داده‌های نمونه‌ای می‌باشد.

بنابراین:

$$e(\bar{X}_n, \bar{X}_n) \approx 4\sigma^2 [f(\bar{\mu})]^2$$

بنابراین دلیل دیگر برای صحت ادعای بیان شده، با توجه به رابطه $e(\bar{X}_n, \bar{X}_n) \approx 4\sigma^2 [f(\bar{\mu})]^2$ روش می‌گردد. از آنجا که $[f(\bar{\mu})]^2$ اصولاً ثابت می‌باشد، لذا هر چه σ^2 کمتر باشد، دنباله‌ها (دمها) ای توزیع کوتاه‌تر و هر چه σ^2 بیشتر باشد، دنباله‌های توزیع طولانی‌تر و در نتیجه کارآیی میانه بالاتر خواهد بود. همچنین با توجه به اینکه میانگین حسابی به شدت تحت تأثیر داده‌ها قرار دارد، لذا هر چه پراکندگی در داده‌ها بیشتر باشد، نوسانات میانگین شدیدتر خواهد شد، حال آنکه در میانه زیاد محسوس نمی‌باشد. از این‌رو در چنین وضعیتها بایی میانه بر میانگین ترجیح داده می‌شود. دو توزیع زیر، گواه بیشتری برای مدعای هستند.

۷-۳ توزیع کشی بریده شده

فرض کنیم X_1, X_2, \dots, X_n نمونه‌ای تصادفی از چگالی زیر باشند ($d > 0$).

$$f_d(x) = \begin{cases} \frac{\pi}{\tan^{-1}(d) - \tan^{-1}(-d)} f(x) & ; |x| \leq d \\ 0 & ; \text{سایر نقاط} \end{cases}$$

که در آن

$$f(x) = \frac{1}{\pi(1+x^2)} ; x \in R$$

داریم:

$$E_d X = 0 \Rightarrow \mu = \bar{\mu} = 0$$

$$\text{var}(x) = \frac{2d}{\tan^{-1}(d) - \tan^{-1}(-d)} - 1$$

و

$$e(\bar{X}_n, \bar{X}_n) \approx \frac{4[2d - \tan^{-1}(d) + \tan^{-1}(-d)]}{[\tan^{-1}(d) - \tan^{-1}(-d)]^2}$$

بنابراین:

$+\infty$	۵	۴	۳	۱	d
$+\infty$	$1/41$	$1/13$	$0/896$	$0/44$	$e(\bar{X}_n, \bar{X}_n)$

قضیه ۷-۳: برای چگالی‌های متقارن تک نمایی یا بطور کلی تر چگالی‌های متقارن حول صفر که برای هر x در برد متغیر تصادفی X ، $f(x) \leq f(0)$ باشد، داریم:

$$e(\bar{X}_n, \bar{X}_n) \geq \frac{1}{3}$$

تساوی زمانی برقرار می‌باشد که توزیع مورد نظر ما یکنواخت باشد. برای روشن شدن مطلب، به بررسی توزیع یکنواخت می‌پردازیم:

۸-۳ توزیع یکنواخت

فرض کنیم X_1, X_2, \dots, X_n یک نمونه تصادفی از توزیع زیر باشند.

$$f_X(x; \theta) = \frac{1}{\theta} ; 0 \leq x \leq \theta$$

۶-۳ توزیع نرمال بریده شده

فرض کنیم X_1, X_2, \dots, X_n نمونه‌ای تصادفی از توزیع زیر باشند ($\alpha > 0$).

$$f_a(x) = \begin{cases} \frac{f(x)}{\Phi(a)-1} & ; |x| \leq a \\ 0 & ; \text{سایر نقاط} \end{cases}$$

که در آن $(.)\Phi$ تابع توزیع نرمال استاندارد و

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) ; x \in R$$

یک توزیع متقارن تک نمایی حول 0 است. داریم:

$$\mu = \bar{\mu} = 0$$

$$\text{Var}(X) = 1 - \frac{2a}{\sqrt{2\pi}(2\Phi(a) - 1)} \exp\left(-\frac{a^2}{2}\right)$$

بنابراین:

$$e(\bar{X}_n, \bar{X}_n) \approx \frac{2}{\pi(2\Phi(a) - 1)^2} - \frac{4a \exp\left(-\frac{a^2}{2}\right)}{(2\Phi(a) - 1)^2 \pi \sqrt{2\pi}}$$

در نتیجه:

$-\infty$	۳	۲	۱	a
$0/64$	$0/63$	$0/55$	$0/397$	$e(\bar{X}_n, \bar{X}_n)$

اگر فرضًا قرار دهیم $\mu = \nu - 1$, خواهیم داشت:

$$var(\bar{X}_n) \approx \frac{1}{4n[f(\nu - 1)]^2}$$

در نتیجه با توجه به:

$$e(\bar{X}_n, \bar{X}_n) \approx 4\nu[f(\nu - 1)]^2$$

داریم:

۲۰	۱۰	۵	۴	۳	۲	ν
۰/۶۹	۰/۷۲	۰/۷۸	۰/۸۲۰	۰/۸۹۶	۱/۰۳	$e(\bar{X}_n, \bar{X}_n)$

که بیانگر همان حقیقت گرایش توزیع خی دو به توزیع نرمال در حد (وقتی $\nu \rightarrow \infty$) است. به ازاء $\nu = 2, 3 = 7$ برتری میانه بر میانگین نشان داده شده است. تقریباً هر مقداری در بازه $[\mu, \mu]$ اختیار نماید، رفتار کارآیی مشابه جدول فوق خواهد بود.

۵ نتیجه

به طور کلی، می‌توان چنین استنباط نمود که به خصوص در توزیعهای متقارن تک نمایی بدون مقدمه و صرفنظر از ساختار چگالی توزیع، نمی‌توان میانه یا میانگین را برآوردگری m.co. و کارا جهت مرکزیت توزیع دانست. این موضوع حتی در توزیعهای نامتقارن هم مشهود است و دقیقاً به ساختار و چگونگی چگالی بستگی دارد.

داریم:

$$\mu = \bar{\mu} = \frac{\theta}{2}$$

$$var(\bar{X}_n) = \frac{\theta^2}{12n}, var(\bar{X}_n) \simeq \frac{\theta^2}{4n}$$
و در نتیجه:

$$e(\bar{X}_n, \bar{X}_n) \simeq \frac{1}{3} < 1$$
تبصره: اگر واریانس دقیق \bar{X}_n را نیز بدست بیاوریم، باز هم کارآیی $\frac{1}{3}$ می‌گردد.

۹-۳ توزیع خی دو

فرض کنیم X_1, X_2, \dots, X_n نمونه‌ای تصادفی از چگالی نامتقارن زیر باشد.

$$f_X(x) = \frac{1}{\Gamma(\frac{\nu}{2}) 2^{\frac{\nu}{2}}} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}, \quad x \geq 0$$

داریم:

$$\mu = \nu, \quad var(\bar{X}_n) = \frac{2\nu}{n}$$

همچنین^۸:

$$\bar{\mu} = \nu - 2$$

از طرفی در توزیع خی دو داریم:

$$\nu - 2 \leq \bar{\mu} \leq \nu$$

مراجع

- [1] DUDEWICZ, E.J. & MISHRA (1988), Modern mathematical Statistics, John Wiley & Sons.
- [2] ROHATGI, V.K.,(1976), Introduction to probability theory and Mathematical Statistics, John Wiley & Sons.
- [3] LEHMMAN, E.L.,(1983), Theory of point estimation, John Wiley & Sons
- [4] RAO, C.R. (1973), Linear Statistical Inference and Its applications, John Wiley & Sons.

[۵] برنارد و. لیندگرن، نظریه آمار، جلد اول (ترجمه دکتر ابوالقاسم بزرگ نیا) سال انتشار ۱۹۷۶.

^۸ نمای توزیع خی دو با درجه آزادی می‌باشد.