به دست آوردن برآوردگرهای گشتاوری با استفاده از تابع توزیع تجربی

بهزاد منصوری [۱]، رحیم چینی پرداز[۲] سامی عطیه سید الفرطوسی[۳]و حبیب اله ممبینی[۴]

تاریخ دریافت: ۱۴۰۱/۰۵/۰۸
تاریخ پذیرش: ۱۴۰۱/۱۲/۰۲

چکیده:
از تابع توزیع تجربی به‌عنوان برآورد تابع توزیع احتمال تجمعی یک متغیر تصادفی استفاده می‌شود. تابع توزیع تجربی نقشی اساسی در بسیاری از استنباط‌های آماری دارد که در برخی موارد کمتر شناخته‌شده هستند. در این مقاله تابع احتمال تجربی به‌عنوان مشتق تابع توزیع تجربی معرفی‌شده و با استفاده از آن نشان داده می‌شود که برآوردگرهای گشتاوری مانند میانگین نمونه، میانه نمونه، واریانس نمونه‌ای و ضریب همبستگی نمونه‌ای حاصل جایگزین کردن تابع چگالی متغیر تصادفی با تابع احتمال تجربی آن در تعریف‌های نظری هستند. علاوه بر این، برای برآورد پارامترهای جامعه از برآوردگر هسته تابع چگالی احتمال استفاده شده و یک روش جدید برای برآورد پهنای باند در برآوردگر هسته تابع چگالی احتمال، معرفی‌شده است.

واژه‌های کلیدی: تابع توزیع تجربی، برآوردگر هسته، برآورد گشتاوری، پهنای باند

۱ مقدمه

در علم آمار، برای احتمال رخداد یک پیشامد از یک نمونه تصادفی از فرآیند مولد آن پیشامد، کمک می‌گیرند و فراوانی نسبی حالت¬های مساعد در نمونه را برآوردی از احتمال واقعی، تلقی می‌کنند. در نظریه احتمال مقادیر اندازه گیری شده از نمونه تصادفی را اندازه‌گیری‌های تجربی و فراوانی نسبی آن‌ها را «احتمال تجربی» پیشامد میگویند. برای مثال در پزشکی CRP [۵] از جمله پروتئین‌هایی است که در هنگام وجود التهاب در بدن افزایش می‌یابد. از CRP می‌توان برای پی بردن به ریسک خطر ابتلا به بیماری‌های قلبی استفاده کرد. اگر مقدار CRP در خون بیشتر از ۳ باشد، پزشکان خطر ابتلا به بیماری قلبی عروقی را بسیار بالا ارزیابی می‌کنند. حال فرض کنید که علاقه‌مند هستیم نسبت افرادی که در یک جامعه مشخص کمتر در معرض بیماری‌های قلبی و عروقی هستند را برآورد کنیم. برای این منظور از CRP اندازه‌گیری شده در یک نمونه تصادفی از این جامعه استفاده می‌کنیم. با فرض اینکه حجم نمونه تصادفی n باشد، احتمال تجربی پایین بودن ریسک خطر حمله قلبی (p) به‌صورت زیر محاسبه می‌شود:

$$\hat{p} = \frac{1}{n}\{\text{تعدادی از } n \text{ نفر که } CRP \text{ آن‌ها کمتر یا مساوی ۳ است}\}.$$

به‌سادگی می‌توان نشان داد که $\hat{p} \xrightarrow{P} p$ هرگاه $n \to \infty$. در اینجا $\xrightarrow{P}$ نماد همگرایی در احتمال است.

در مثال بالا فرض کنید که متغیر تصادفی X نشان‌دهنده میزان CRP برای یک فرد معین در این جامعه باشد. در این صورت $p = P(X \le 3) = F_X(3)$ است که در آن $F_X(x)$ تابع توزیع تجمعی متغیر تصادفی X است. مفهوم احتمال تجربی ارتباطی نزدیک با تابع توزیع تجربی دارد. برای نمونه تصادفی مستقل و هم‌توزیع $X_1, \ldots, X_n$ تابع توزیع تجربی به‌صورت

$$F_n(X) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

تعریف می‌شود که در آن $I(x)$ تابع نشانگر است. در مثال بالا داریم:

$$F_n(3) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le 3) = \hat{p}.$$

[۱] گروه آمار- دانشکده علوم ریاضی و کامپیوتر- دانشگاه شهید چمران اهواز (نویسنده مسئول: b.mansouri@scu.ac.ir)
[۲] گروه آمار- دانشکده علوم ریاضی و کامپیوتر- دانشگاه شهید چمران اهواز
[۳] گروه آمار- دانشکده علوم ریاضی و کامپیوتر- دانشگاه شهید چمران اهواز
[۴] گروه آمار- دانشکده علوم ریاضی و کامپیوتر- دانشگاه شهید چمران اهواز

[5] C-Reactive Protein

به همین دلیل تابع توزیع تجربی را برآوردگر فراوانی نسبی نیز میگویند. سادگی و نرخ همگرایی سریع تابع توزیع تجربی سبب رواج استفاده از آن شده است به‌گونه‌ای که در مباحث آمار ناپارامتری به‌صورت گسترده‌ای از تابع توزیع تجربی به‌عنوان برآورد تابع توزیع مجهول یک متغیر تصادفی استفاده می‌شود. در واقع بسیاری از استنباط‌های آماری بر پایه تابع توزیع تجربی هستند. برای مثال فلسفه روش بوت استرپ ناپارامتری در استفاده از نمونه‌گیری با جایگذاری برای بازنمونه‌گیری (دادن شانس برابر به مشاهدات) بر مبنای استفاده از تابع توزیع تجربی برای متغیر تصادفی تحت مطالعه است. از کاربردهای متنوع دیگر تابع توزیع تجربی، می‌توان به آزمون کولموگروف- اسمیرنوف در آزمون‌های هم توزیعی، برآورد تابع مفصل و برآورد منحنی عملکرد گیرنده (ROC)[۶] اشاره کرد. در استنباط آماری، یک روش رایج برآورد پارامترهای مجهول استفاده از روش گشتاوری است. علیرغم اینکه روش گشتاوری برآورد پارامتر، در کنار روش حداکثر درستنمایی از رایج‌ترین روش‌ها است، به‌ندرت به مبانی نظری آن توجه شده است. [۱۰] با تعریف برآورد جایگزین [۷] به رابطه بین تابع توزیع تجربی و برآورد گشتاوری اشاره‌کرده است.

در این مقاله هدف نشان دادن اهمیت تابع توزیع تجربی و نقش آن در به دست آوردن برآوردهای گشتاوری به‌صورت روشن و ساده است. برای این منظور با استفاده از تابع توزیع تجربی و تابع دلتای دیراک، تابع احتمال تجربی را تعریف کرده و نشان می‌دهیم که برآوردگرهای گشتاوری پارامترهای جامعه از جایگزین کردن تابع احتمال تجربی به‌جای تابع چگالی احتمال در تعریف‌های نظری به دست می‌آیند. علاوه بر این از برآوردگر هسته تابع چگالی احتمال برای برآورد پارامترهایی مانند میانگین و واریانس استفاده کرده و یک روش جدید برای برآورد پهنای باند پیشنهاد می‌کنیم. مقاله در پنج بخش تنظیم‌شده است. در بخش دوم برخی از خواص تابع توزیع تجربی به‌اختصار مرور شده است. در بخش سوم تابع احتمال تجربی معرفی‌شده است. در این بخش رابطه بین تابع توزیع تجربی و تابع احتمال تجربی با برآوردگرهای گشتاوری نشان داده شده است. در بخش چهارم مقاله با استفاده از برآوردگر هسته تابع چگالی احتمال، برآوردگرهایی برای میانگین و واریانس جامعه به‌دست‌آمده است. بخش پنجم مقاله به بحث و نتیجه‌گیری اختصاص‌یافته است.

۲ تابع توزیع تجربی

فرض کنید که $X_1, \ldots, X_n$ یک دنباله از متغیرهای تصادفی مستقل و هم‌توزیع با توزیع $F(x)$ باشد. با توجه به هم‌توزیع و مستقل بودن X_i ها که دارای توزیع برنولی هستند، $nF_n(x)$ دارای توزیع دوجمله‌ای با پارامترهای n و $p = F(x)$ است؛ بنابراین

$$E(F_n(x)) = F(x), Var(F_n(x)) = \frac{F(x)(1 - F(x))}{n}$$

در نتیجه $MSE(F_n(x)) = n^{-1}F(x)(1-F(x))$. بنابراین هرگاه $n \to \infty$ آنگاه $F_n(x) \xrightarrow{p} F(x)$. مرتبه همگرایی تابع توزیع تجربی $O_P(n^{-1/2})$ است یا به عبارتی با نرخ $\sqrt{n}$ سازگار است زیرا برای $\epsilon > 0$ دلخواه و ثابت متناهی M به‌صورت $M = \frac{F(x)(1-F(x))}{\epsilon}$ از نامساوی چیبیشف داریم:

$$P(|\sqrt{n}(F_n(x) - F(x))| > M^{1/2}) \leq \frac{nE(F_n(x) - F(x))^2}{M} = \epsilon.$$

می‌توان با استفاده از قانون قوی اعداد بزرگ نشان داد که برای هر x ثابت $F_n(x)$ با احتمال یک به $F(x)$ همگرا است. به‌عبارت‌دیگر هرگاه $n \to \infty$ آنگاه با احتمال یک داریم:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \to E[I(X_i \leq x)] = F(x) \qquad (1)$$

این نتیجه ثابت می‌کند که $F_n(x)$ به‌صورت نقطه‌به‌نقطه [۸] به $F(x)$ همگرا است . [۳] ، [۱] مستقلاً در سال ۱۹۳۳ ثابت کردند که $F_n(x)$ به‌صورت یکنواخت نیز به $F(x)$ همگرا است؛ یعنی هرگاه $n \to \infty$ آنگاه با احتمال یک داریم:

$$\sup_x |F_n(x) - F(x)| \to 0.$$

دقت کنید که $\sup_x |F_n(x) - F(x)|$ آماره کولموگروف- اسمیرنوف برای آزمون هم توزیعی است. علاوه بر این هرگاه $n \to \infty$ از قضیه حد مرکزی داریم:

$$\sup_x \sqrt{n}|F_n(x) - F(x)| \sim N(0, F(x)(1 - F(x))). \qquad (2)$$

با استفاده از قانون قوی اعداد بزرگ $F_n(x) - F(x)$ همگرا است درحالی‌که طبق قضیه حد مرکزی $\sqrt{n}(F_n((x) - F(x))$ واگرا است (دامنه متغیر نرمال استاندارد کل اعداد حقیقی است). به بیان ساده‌تر $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i \leq x)$ همگرا است اما $\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$ واگرا است. حال سؤال این است که مرز همگرایی و واگرایی کجاست؟ به‌عبارت‌دیگر تا چه توانی از n عبارت $(F_n(x) - F(x))$ همگرا است؟

[6] Receiver operating curve

[7] Plug- in estimator

[8] Pointwise convergence

قانون لگاریتم تکراری این مرز را مشخص می‌کند. بر طبق قانون لگاریتم تکراری [۹]

$$\lim_{n\to\infty} \frac{\sqrt{n}\sup_x |F_n(x)-F(x)|}{\sqrt{۲\ln\ln n}} \le \frac{۱}{۲}.$$

برای توضیح بیشتر درباره اثر ضرایب $\frac{۱}{n}$، $\frac{۱}{\sqrt{n}}$ و $\frac{۱}{\sqrt{۲n\ln\ln n}}$ در شکل (۱) نمودار این ضرایب به‌عنوان توابعی از n رسم شده و سرعت همگرایی آن‌ها به صفر را نمایش داده شده است. دقت کنید که چگونه منحنی $\frac{۱}{\sqrt{۲n\ln\ln n}}$ مقادیری بین منحنی $\frac{۱}{n}$ و $\frac{۱}{\sqrt{n}}$ قرار می‌گیرد. خینچین (۱۹۲۴) [۱۰] و کولموگروف (۱۹۲۹) [۱۱] قانون لگاریتم تکراری را معرفی و مطالعه کرده‌اند. برای مطالعه بیشتر در این زمینه [۴] فصل دوم را ببینید.

۱.۲ فاصله اطمینان برای تابع توزیع تجمعی

از رابطه (۲) نمی‌توان یک فاصله اطمینان کاربردی برای تابع توزیع مجهول $F(x)$ به دست آورد زیرا حدود فاصله اطمینان به $F(x)$ وابسته خواهند بود. [۲] نشان دادند که برای هر $\epsilon > ۰$

$$P(\sup_x |F_n(x)-F(x)| > \epsilon) \le ۲e^{-۲n\epsilon^۲}. \qquad (۳)$$

این نامساوی که به نامساوی دوورتسکی-کیفر-ولفویتز (DKW) معروف است امکان ساختن یک فاصله اطمینان کاربردی برای $F(x)$، را فراهم می‌کند. برای این منظور کافی است که برای $\epsilon_n = \sqrt{\frac{۱}{۲n}\ln(\frac{۲}{\alpha})}$ تعریف کنیم

$$L(x)=\max\{F_n(x)-\epsilon_n, ۰\},\ U(x)=\max\{F_n(x)+\epsilon_n, ۱\}, \quad (۴)$$

آنگاه از رابطه (۳) داریم:

$$P(L(x)\le F(x)\le U(x)) \ge ۱-\alpha.$$

با استفاده از نظریه [۹] می‌توان نامساوی‌های کلی‌تری را به دست آورد ([۱۰]). [۱۲] در حالت یک‌بعدی، با استفاده از قضیه وپنیک و چروننکیس (VC) می‌توان نشان داد که

$$P(\sup_x |F_n(x)-F(x)| > \epsilon) \le ۸(n+۱)e^{-\frac{n\epsilon^۲}{۳۲}}. \qquad (۵)$$

(برای جزئیات بیشتر [۱۰] را ببینید). با استفاده از (۵) و به ازای $\epsilon_n = \sqrt{\frac{۳۲}{n}\ln(\frac{۸(n+۱)}{\alpha})}$ در حدود رابطه (۴) یک فاصله اطمینان در سطح $۱-\alpha$ برای $F(x)$ با استفاده از نامساوی VC به دست می‌آید اما دامنه این فاصله اطمینان نسبت به فاصله اطمینانی که با استفاده از نامساوی DKW به دست آمد، بسیار بزرگ‌تر بوده و کارا نیست. در ادامه این دو فاصله اطمینان را برای یک مجموعه داده شبیه‌سازی‌شده مقایسه می‌کنیم.

داده‌های این قسمت یک مجموعه آمیخته از ۱۰۰۰ داده شبیه‌سازی‌شده از توزیع نرمال با میانگین ۱۰ و انحراف معیار ۱ و ۱۰۰۰ داده شبیه‌سازی‌شده از توزیع نرمال با میانگین ۱۵ و انحراف معیار ۱/۵ هستند. شکل (۲) چگالی این مجموعه داده و شکل (۳) تابع توزیع تجربی داده‌ها را به همراه فاصله اطمینان به‌دست‌آمده از نامساوی‌های (DKW به رنگ قرمز) و (VC به رنگ آبی) را نشان می‌دهد.

۳ تابع احتمال تجربی

در این بخش می‌خواهیم از تابع توزیع تجربی برای محاسبه احتمال در هر نقطه از دامنه متغیر تصادفی استفاده کنیم. برای این منظور از تابع دلتای دیراک استفاده کنیم. تابع دلتای دیراک را با $\delta(x)$ نشان می‌دهیم. این تابع در خواص زیر صدق می‌کند:

$$\delta(x)=\begin{cases} ۰ & x\neq ۰,\\ \infty & x=۰ \end{cases}$$

و $\int_{-\infty}^{\infty}\delta(x)dx = ۱$. دقت کنید که

$$\int_{-\infty}^{\infty}\delta(x-a)g(x)dx = \int_{-\infty}^{\infty}\delta(x-a)[g(a)+(g(x)-g(a))]dx = g(a),$$

زیرا $\delta(x-a)(g(x)-g(a)) = ۰$. حال تابع نشانگر را در نظر بگیرید:

$$I(x\ge ۰)=\begin{cases} ۱ & x\ge ۰\\ ۰ & x<۰, \end{cases}$$

مشتق این تابع همه‌جا به‌جز در نقطه صفر برابر صفر است و چون در نقطه صفر ناپیوسته است مشتق‌پذیر نبوده و می‌توان مشتق آن را بینهایت دانست. به همین دلیل [۶] تابع دلتای دیراک را مشتق تابع نشانگر تشخیص دادند یعنی $\delta(x) = \frac{dI(x\ge ۰)}{dx}$. بنابراین اگر تابع احتمال تجربی را به‌عنوان مشتق تابع توزیع تجربی تعریف کرده و با

[9] Law of the iterated logarithm
[10] Khinchin
[11] Kolmogorov
[12] Wasserman

$f_n(x)$ نشان دهیم، آنگاه داریم:

$$f_n(x) := \frac{dF_n(x)}{dx} = \frac{1}{n}\sum_{i=1}^{n}\delta(X_i - x).$$

در واقع تابع احتمال تجربی در همسایگی بسیار کوچک هر مشاهده وزن $\frac{1}{n}$ را اعمال می‌کند. باوجوداینکه تابع احتمال تجربی به معنی دقیق یک تابع احتمال نیست، اما چنانکه در ادامه خواهیم دید، استفاده از آن برای نشان دادن فلسفه نظری برآوردهای گشتاوری و رابطه آن‌ها با تابع توزیع تجربی بسیار مفید است. با توجه به مستقل و هم‌توزیع بودن مشاهدات و تعریف تابع دلتای دیراک داریم:

$$E((f_n(x)) = E(\delta(Y - x)) = \int_{-\infty}^{\infty}\delta(Y - x)f(y)dy = f(x)$$

و

$$\begin{aligned} Var(f_n(x)) &= \frac{1}{n}Var(\delta(Y - x)) \\ &= \frac{1}{n}\{E(\delta(Y - x)^2) - (E(\delta(Y - x)))^2\} \\ &= \frac{f(x)(1 - f(x))}{n}. \end{aligned}$$

بنابراین هرگاه $n \to \infty$ آنگاه $f_n(x) \xrightarrow{p} f(x)$.

اگر امید ریاضی تجربی یک تابع مانند $g(x)$ را با $E_n(g(X))$ نشان دهیم، با توجه به تعریف تابع دلتای دیراک داریم:

$$\begin{aligned} E_n(g(X)) &= \int_{-\infty}^{\infty} g(x)dF_n(x) \\ &= \int_{-\infty}^{\infty} g(x)f_n(x)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} g(x)\delta(X_i - x)dx \\ &= \frac{1}{n}\sum_{i=1}^{n} g(X_i). \end{aligned}$$

برای مثال

$$E_n(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, \quad E_n(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \bar{X^2}$$

و

$$\sigma_n^2(X) = E_n(X^2) - (E_n(X))^2 = \bar{X^2} - \bar{X}^2.$$

که عبارت فوق برآوردگر گشتاوری واریانس جامعه است.

۱.۳ تابع احتمال تجربی دومتغیره

برای نمونه تصادفی مستقل و هم‌توزیع $(X_1, Y_1), \ldots, (X_n, Y_n)$ با تابع توزیع احتمال توأم $F(x, y)$ و تابع چگالی احتمال توأم $f(x, y)$، تابع توزیع تجربی دومتغیره به‌صورت

$$F_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x_i, Y_i \le y_i)$$

تعریف می‌شود. تابع احتمال توأم تجربی دو متغیر تصادفی X و Y را به‌صورت زیر تعریف می‌کنیم،

$$f_n(x, y) := \frac{dF_n(x, y)}{dxdy} = \frac{1}{n}\sum_{i=1}^{n}\delta((X_i - x_i)(Y_i - y_i)).$$

امید ریاضی تجربی یک تابع مانند $g(X, Y)$ عبارت است از

$$\begin{aligned} E_n(g(X, Y)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f_n(x, y)dxdy \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)\delta((X_i - x)(Y_i - y))dxdy \\ &= \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i) \end{aligned}$$

بنابراین

$$E_n(XY) = \frac{1}{n}\sum_{i=1}^{n} X_iY_i = \bar{XY},$$

در نتیجه

$$\sigma_n(xy) = E_n(XY) - E_n(X)E_n(Y) = \bar{XY} - \bar{X}\bar{Y},$$

برآوردگر تجربی کوواریانس دو متغیر تصادفی است و برآوردگر تجربی ضریب همبستگی نیز به‌صورت

$$r_n(x, y) = E_n(XY) = \frac{\sigma_n(x, y)}{\sigma_n(x)\sigma_n(y)} = \frac{\bar{XY} - \bar{X}\bar{Y}}{\sqrt{\bar{X^2} - \bar{X}^2}\sqrt{\bar{Y^2} - \bar{Y}^2}}$$

است؛ یعنی ضریب همبستگی نمونه‌های نیز حاصل جایگزین کردن توابع چگالی مجهول $f(y)$, $f(x)$ و $f(x, y)$ با معادل‌های تجربی آن‌ها در تعریف نظری ضریب همبستگی است.

۲.۳ میانه تجربی

برای نمونه تصادفی مستقل و هم‌توزیع $X_1, \ldots, X_n$ فرض کنید که $X_{(i)}, i = 1, \ldots, n$ نشان‌دهنده i امین آماره ترتیبی در نمونه باشد. در این صورت اگر n فرد باشد آنگاه با تعریف

$$P_n(X \le a) := F_n(a) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le a),$$

داریم:

$$P_n\Big(X \le X_{(\frac{n+1}{2})}\Big) = P_n\Big(X \ge X_{(\frac{n+1}{2})}\Big),$$

یعنی $X_{(\frac{n+1}{2})}$ میانه نمونه با استفاده از تابع توزیع تجربی داده‌ها است. اگر n زوج باشد آنگاه با تعریف

$$Med = \frac{1}{2}\Big[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\Big],$$

داریم:

$$P_n(X \le Med) = P_n(X \ge Med).$$

باید دقت کرد که همه برآوردگرهای نمونه‌ای حاصل جایگزین کردن تابع توزیع تجربی به‌جای تابع توزیع مجهول در روابط نظری نیستند. در بخش بعد نشان می‌دهیم که برخی از برآوردگرهای نااریب پارامترهای جامعه حاصل جایگزین کردن تابع توزیع مجهول با برآوردگر هسته هستند.

۴ برآوردگر چگالی هسته

در بخش قبل برآوردگرهای پارامترهای جامعه را با جایگزین کردن تابع احتمال تجربی به‌جای تابع احتمال مجهول به دست آوردیم. در این بخش می‌خواهیم برآوردگرها را با جایگزین کردن برآوردگر هسته تابع چگالی به‌جای تابع احتمال مجهول در روابط نظری به دست آورده و با نتایج بخش قبل مقایسه کنیم.

برای نمونه تصادفی مستقل و هم‌توزیع $X_1, \ldots, X_n$ با تابع چگالی احتمال $f(x)$، برآوردگر تابع چگالی احتمال به فرم

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\Big(\frac{x - X_i}{h}\Big),$$

تعریف می‌شود که در آن $h = h_n$ پهنای باند است و در شرایط $\lim_{n\to\infty} h = 0$ و $\lim_{n\to\infty} nh = \infty$ صدق می‌کند و تابع هسته $k(x)$ نیز در شرایط

$$\int_{-\infty}^{\infty} k(x)dx = 1, \quad \int_{-\infty}^{\infty} xk(x)dx = 0, \quad \int_{-\infty}^{\infty} x^2 k(x)dx = \mu_2(k) \le \infty, \quad (۶)$$

صدق می‌کند. تحت این فرض‌ها و با فرض اینکه مشتق دوم تابع $f(x)$ انتگرال‌پذیر است می‌توان نشان داد که $\hat{f}(x)$ در احتمال به $f(x)$ همگرا است ([۸] صفحات ۴۰-۳۹ را ببینید.) همچنین می‌توان نشان داد که $\hat{f(x)}$ با احتمال یک نیز به $f(x)$ همگرا است ([۶] قضیه ۱-۴ و ۱-۵).

۱.۴ امید ریاضی تابعی از یک متغیر تصادفی پیوسته با استفاده از برآوردگر چگالی هسته

فرض کنید $g(x)$ که تابعی از متغیر تصادفی پیوسته X با تابع چگالی احتمال $f(x)$ بوده و $\hat{f(x)}$ برآوردگر هسته تابع چگالی با استفاده از نمونه تصادفی $X_1, \ldots, X_n$ باشد. اگر $E_{\hat{f}}$ نشان‌دهنده امید ریاضی با جایگزینی $\hat{f}(x)$ به‌جای $f(x)$ باشد و با فرض اینکه تابع هسته در روابط (۶) صدق کند داریم:

$$\begin{aligned} E_{\hat{f}}(g(X)) &= \int_{-\infty}^{\infty} g(x)\hat{f}(x)dx \\ &= \frac{1}{nh}\int_{-\infty}^{\infty} g(x)\sum_{i=1}^{n} k\Big(\frac{x - X_i}{h}\Big)dx \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{\infty} g(x)k\Big(\frac{x - X_i}{h}\Big)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} g(t.h + X_i)k(t)dt \\ &= \frac{1}{nh}\sum_{i=1}^{n} E_k(g(t.h + X_i)). \end{aligned}$$

برای مثال

$$\begin{aligned} E_{\hat{f}}(X) &= \frac{1}{n}\sum_{i=1}^{n} E_k(t.h + X_i) \\ &= \frac{1}{n}\sum_{i=1}\int_{-\infty}^{\infty}(t.h + X_i)k(t)dt = \bar{X}, \end{aligned}$$

زیرا طبق شرایط (۶) داریم: $\int xk(x)dx = 0$، $\int k(x)dx = 1$. همچنین

$$\begin{aligned} E_{\hat{f}}(X^2) &= \frac{1}{n}\sum_{i=1}^{n} E_k(t.h + X_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(t^2h^2 + 2t.hX_i + X_i^2)k(t)dt \\ &= h * 2\int_{-\infty}^{\infty} t^2 k(t)dt + \bar{X}^2, \end{aligned}$$

در این صورت:

$$\begin{aligned} Var_{\hat{f}}(\bar{X}) &= E_{\hat{f}}(X^2) - (E_{\hat{f}}(X))^2 \\ &= \bar{X}^2 - \bar{X}^2 + h^2\mu_2(k), \end{aligned}$$

که در آن $\mu_2(k) = \int t^2 k(t)dt$ واریانس تابع هسته است. فرض کنید X_i ها نمونه‌ای تصادفی از توزیع $f(x)$ با میانگین μ و واریانس σ^2 باشند. آنگاه با انتخاب $\tilde{h} = \sqrt{\frac{\sigma^2}{n\mu_2(k)}}$ داریم:

$$\begin{aligned} Var_{\hat{f}_{\tilde{h}}}(X) &= E_{\hat{f}_{\tilde{h}}}(X^2) - (E_{\hat{f}_{\tilde{h}}})^2 \\ &= \bar{X}^2 - \bar{X}^2 + \frac{\sigma^2}{n}, \end{aligned}$$

و میدانیم که

$$E\Big(\bar{X}^2 - \bar{X}^2 + \frac{\sigma^2}{n}\Big) = \sigma^2,$$

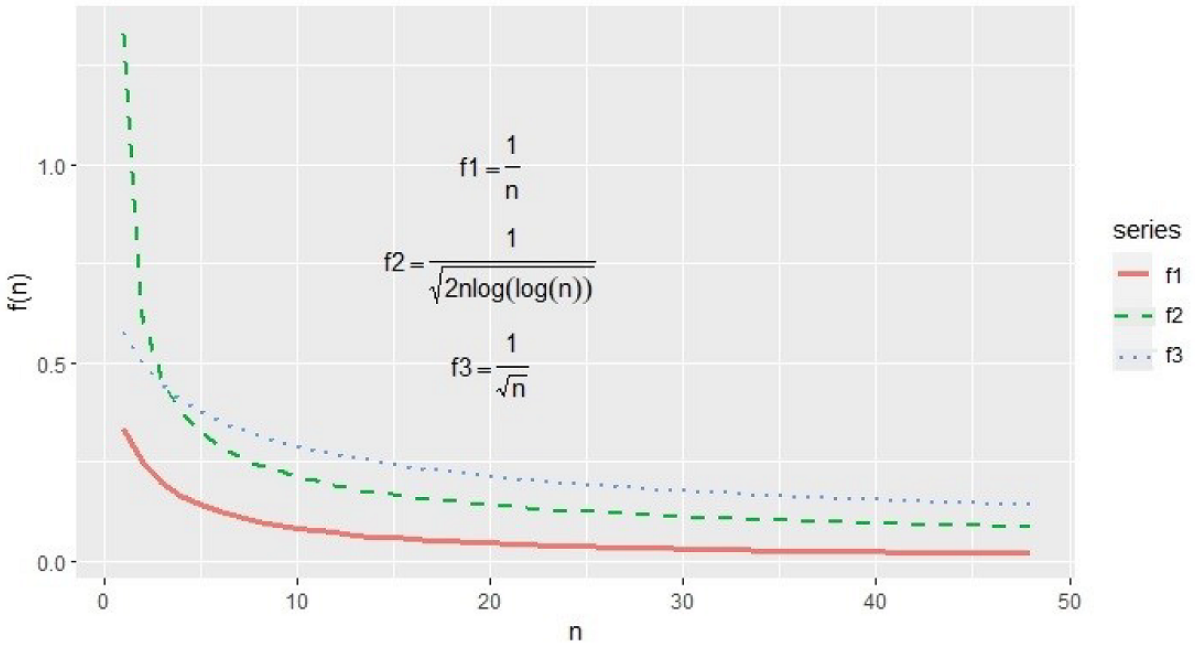
یعنی با انتخاب درست پهنای باند، $Var_{\hat{f}_{\tilde{h}}}(X)$ برآوردگر واریانس با استفاده از تابع چگالی هسته، برآوردگر نااریب واریانس جامعه است درحالی‌که همان‌طور که در بخش قبل دیدیم σ_n^2 (برآوردگر واریانس با

استفاده از تابع توزیع تجربی) اریب است. دقت کنید که به ازای تابع هسته نرمال استاندارد $\mu_۲(k) = ۱$ است و در نتیجه $\tilde{h} = \sigma n^{-\frac{۱}{۲}}$ به دست می‌آید.

[۸] با یک قاعده سرانگشتی [۱۳] برآوردگر زیر را برای پهنای باند به دست آورد:

$$h_{ROT.f} = ۱/۰۶\sigma n^{-\frac{۱}{۵}}$$

سیلورمن (صفحات ۴۸- ۴۵) این برآوردگر را نقد و ارزیابی کرده و نتیجه گرفت که اگر چگالی واقعی داده‌ها با توزیع نرمال متفاوت باشد مثلاً چند مدی، چوله یا برجسته باشد، کارایی این برآوردگر پایین بوده و سبب بیش هموواری در چگالی برآورد شده می‌گردد. با توجه به اینکه $\tilde{h} < h_{ROT.f}$ انتظار این است که $h_{ROT.f}$ نسبت به $\tilde{h}$ سبب بیش هموواری در چگالی برآورد شده، شود. با توجه به‌سادگی $\tilde{h}$ مطالعه بیشتر آن و مقایسه عملکرد آن با دیگر برآوردگرهای پهنای باند می‌تواند موضوع جالبی برای تحقیق باشد.

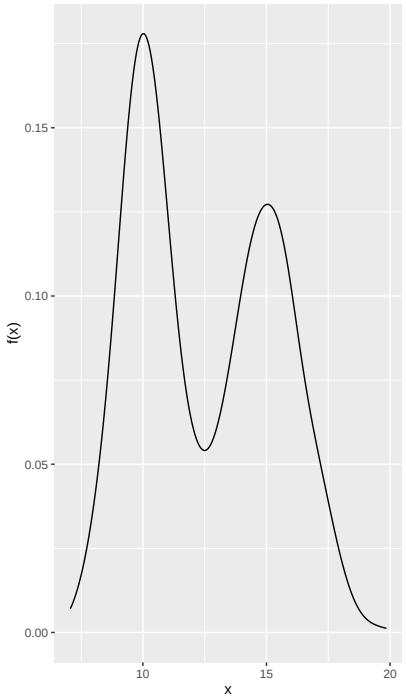


شکل ۱. مقایسه سرعت همگرایی دنباله $\frac{۱}{n}$, $\frac{۱}{\sqrt{n}}$ و $\frac{۱}{\sqrt{۲n \ln \ln n}}$ به ازای مقادیر n در فاصله ۳ تا ۵۰.

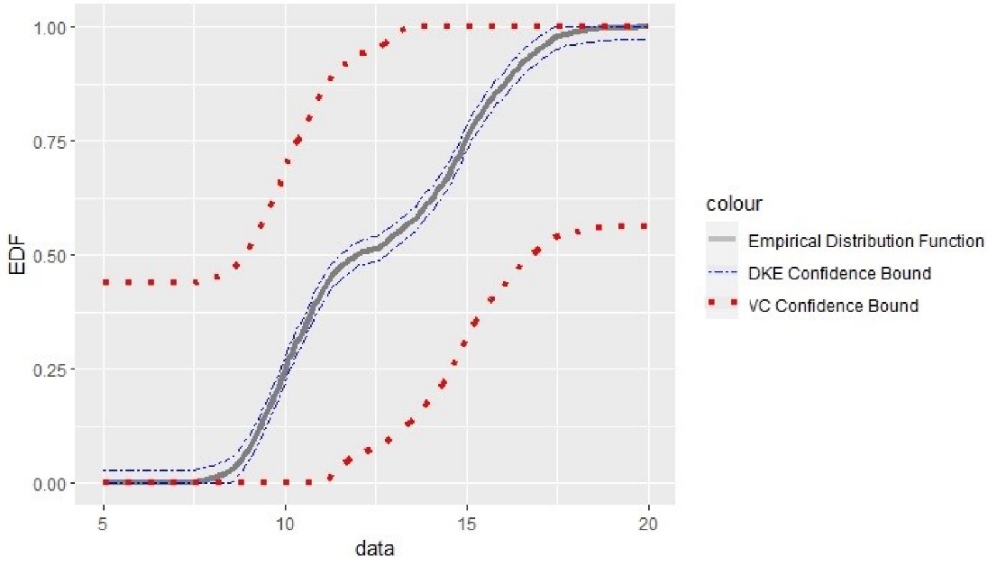
۵ بحث و نتیجه‌گیری

تابع توزیع تجربی، در عین سادگی نقشی پایه‌ای در آمار استنباطی به‌خصوص آمار ناپارامتری دارد. در این مقاله با استفاده از تابع توزیع تجربی و تابع دلتای دیراک، تابع احتمال تجربی را تعریف کرده و نشان دادیم که برآوردگرهای گشتاوری از جایگزین کردن تابع چگالی مجهول با تابع احتمال تجربی، در تعریف‌های نظری به دست می‌آیند. ضعف تابع توزیع تجربی غیر هموار بودن آن است. به همین دلیل برآوردگرهای نوع هسته، رواج یافته‌اند. با این وجود برآوردگرهای نوع هسته نیز در برخی موارد تابع توزیع تجربی استفاده می‌کنند. برای مثال، در برآورد تابع توزیع احتمال به روش هسته، از تابع توزیع تجربی به‌عنوان یک محک استفاده می‌شود (به‌عنوان یک نمونه، [۷] را ببینید) باوجود قدمت تابع توزیع تجربی، تحقیق در مورد آن ادامه دارد ([۵]).

[13] Rule-of-thumb



شکل ۲. تابع چگالی نرمال آمیخته



شکل ۳. تابع توزیع تجربی داده‌ها (منحنی پیوسته) به همراه فاصله اطمینان به‌دست‌آمده از نامساوی‌های (DKW منحنی خط-نقطه چین به رنگ آبی) و (VC منحنی نقطه‌چین به رنگ قرمز).

مراجع

[1] Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, **4**, 421-424.

[2] Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, **27(3)**, 642-669.

[3] Glivenko, V. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, **4**, 92-99.

[4] Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

[5] Langrené, N., and Warin, X. (2021). Fast multivariate empirical cumulative distribution function with connection to kernel density estimation. *Computational Statistics and Data Analysis*, **162**, 107267.

[6] Li, Q., and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

[7] Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions. *Journal of Statistical Planning and Inference*, **35(1)**, 65-75.

[8] Silverman, B. W. (1986). *Density Estimation for Statistics and Data*. Chapman and Hall: London

[9] Vapnik, V. N., and Chervonenkis, A. Y. (1971). Theory of uniform convergence of frequencie of appearance of attributes to their probabilities and problems of defining optimal solution by empiric data.*Avtomatika i Telemekhanika*, **2**, 42-53.

[10] Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Science and Business Media.13 Sarda.

Obtaining moment estimators using the empirical distribution function

Behzad Mansouri [1], Rahim Chinipardaz[1], Sami Atiyah Sayyid Al-Farttosic[1] and Habiballah Mombeni[1]

Abstract:

Empirical distribution function is used as an estimate of the cumulative probability distribution function of a random variable. The empirical distribution function has a fundamental role in many statistical inferences, which are little known in some cases. In this article, the empirical probability function is introduced as a derivative of the empirical distribution function, and it is shown that moment estimators such as sample mean, sample median, sample variance, and sample correlation coefficient result from replacing the random variable density function with the empirical probability function in the theoretical definitions. In addition, the kernel probability density function estimator is used to estimate the population parameters and a new method for bandwidth estimation in the kernel density estimation is introduced.

Keywords: Empirical distribution function, moment estimate, kernel estimator, bandwidth.

[1] **Department of Statistics, Faculty of Mathematics and Comupter Science, Shahid Chamran University of Ahvaz, Ahvaz, Iran**