



یک تقریب جدید برای توزیع صفر آماری آزمون نسبت درست‌نمایی برای بررسی k مشاهده‌ی دورافتاده در یک نمونه‌ی نرمال

رحیم محمودوند^{†*} و حسین حسینی[‡]

[†]دانشگاه پیام نور مرکز تویسرکان

[‡]بانک مرکزی جمهوری اسلامی ایران و دانشگاه کردیف

چکیده. در این مقاله یک تقریب جدید برای توزیع صفر آماری آزمون نسبت درست‌نمایی برای آزمون k مشاهده‌ی دورافتاده‌ی پایینی یا بالایی در یک نمونه‌ی نرمال معرفی شده است. برای بررسی دقت تقریب جدید، مقادیرهای بحرانی به دست آمده از تقریب جدید با مقادیرهایی که توسط توزیع دقیق برای حالت‌های $k = 1, 2$ به دست آمده‌اند، مقایسه شده‌اند. همچنین نتایج با یک روش تقریبی دیگر—که به وسیله‌ی بارتنت و لوییس (۱۹۹۴) ارائه شده است— برای حالت‌های $k = 3, 4$ مقایسه شده‌اند.

واژگان کلیدی. مشاهده‌ی دورافتاده؛ توزیع نرمال؛ آزمون نسبت درست‌نمایی؛ تقریب.

۱ مقدمه

برای انجام دادن آزمون‌های آماری یا روش‌های برآوردیابی، معمولاً فرض می‌شود که مشاهده‌های موجود در یک نمونه‌ی تصادفی، مستقل و هم‌توزیع‌اند. در برخی موارد، فرضی دیگر—که در آزمون‌های آماری، نقش

* نویسنده‌ی عهده‌دار مکاتبات

مهمی را ایفا می‌کند - به دو فرض پیشین افزوده شده و آن این است که مشاهده‌ها از یک جامعه‌ی نرمال استخراج شده‌اند. گاهی در یک نمونه‌ی موجود، دیده می‌شود که یک مشاهده یا مجموعه‌ای از مشاهده‌ها با بقیه ناسازگار است. به عنوان مثال، فرض کنید مشاهده‌ها بر اساس اندازه مرتب شده‌اند و برخی از آن‌ها بیش از حد معمول از مشاهده‌های دیگر بزرگ‌تر یا کوچک‌تر باشند، یا از الگوی توزیع مشاهده‌ها تبعیت نکنند. چنین مشاهده‌هایی، معمولاً دورافتاده نامیده می‌شوند. در این مورد، تعریف‌های مختلفی ارائه شده است (به عنوان مثال، هوکینز (۱۹۸۰)، روسو و لی‌روآ (۱۹۸۷)، بارنت و لوییس (۱۹۹۴)، مور و مکاب (۱۹۹۹) و آگاروال و یو (۲۰۰۱) را ببینید).

یکی از مراحل مهم در هر تحلیل آماری (پیش از آغاز تحلیل)، بررسی وجود مشاهده‌های دورافتاده است. بسیاری از فن‌های آماری به وجود مشاهده‌ی دورافتاده در نمونه حساس‌اند و در صورت وجود چنین مشاهده‌هایی، قابلیت چندانی ندارند. اما شایان ذکر است که روش‌هایی نیز وجود دارند که کم‌تر تحت تأثیر مشاهده‌های دورافتاده قرار دارند (به عنوان مثال، هامپل و دیگران (۱۹۸۶) را ببینید). مشاهده‌های دورافتاده می‌توانند حاصل از خطا باشند، که در آن صورت باید از مجموعه‌ی مشاهده‌ها کنار گذاشته شوند؛ ولی چنان‌چه شواهدی بر این واقعیت وجود نداشته باشد که غیر عادی بودن مشاهده‌ها (ها) به دلیل وجود خطا است، باید با آن‌ها محتاط عمل کرد و تحلیل‌ها را با وجود چنین مشاهده‌هایی انجام داد، اما با تأثیر کم‌تری نسبت به بقیه‌ی مشاهده‌ها (هوپر، ۱۹۸۱).

به طور خلاصه در خصوص مشاهده‌های دورافتاده می‌توان به سه نکته‌ی کلی اشاره کرد: اول این‌که مشاهده‌های دورافتاده از کجا آمده‌اند، دوم این‌که چرا باید مراقب حضور این مشاهده‌ها بود، و سوم این‌که در صورت وجود مشاهده‌های دورافتاده، باید چگونه عمل کرد. روش‌های بسیار زیادی در مورد شناخت مشاهده‌های دورافتاده در یک نمونه‌ی تصادفی بیان شده است، که در ادامه به چند مورد در حد اختصار اشاره می‌شود.

برای تشخیص مشاهده‌های دورافتاده می‌توان از روش‌های عینی - که یک بینش اولیه در مورد مشاهده‌ها ارائه می‌کنند - استفاده کرد. در این‌گونه روش‌ها معمولاً از روش‌های نموداری (مانند هیستوگرام، جعبه‌ای، احتمال نرمال) یا استاندارد کردن مشاهده‌ها (مقایسه‌ی مقادیر استاندارد شده با چندک‌های توزیع نرمال) استفاده می‌کنند (چمبرز و دیگران، ۱۹۸۳)؛ اما بدیهی است که این‌گونه قضاوت از دقت کافی برخوردار نیست و متکی به سلیقه‌ی تحلیل‌گر در خصوص انتخاب مشاهده‌ی دورافتاده است.

علاوه بر این روش‌ها می‌توان از آزمون‌های آماری نیز برای تشخیص مشاهده‌ی دورافتاده استفاده کرد. از این آزمون‌ها می‌توان برای تشخیص چولگی توزیع مشاهده‌ها در حالتی نیز که اندازه‌ی نمونه کوچک است، استفاده کرد (دیوید، ۱۹۸۱). همان‌طور که پیش‌تر بیان شد، اهمیت وجود توزیع نرمال و استفاده از آن در آزمون‌های آماری بسیار مهم است. در این خصوص، آزمون‌هایی به منظور کشف مشاهده‌های

دورافتاده (هنگامی که نمونه‌ی استخراج شده از توزیع نرمال باشد) به‌وسیله‌ی فرگوسن (۱۹۶۱)، گراس (۱۹۶۹)، جانسون و هانت (۱۹۷۹) و تود و دیگران (۱۹۸۳) ارائه شده است. آزمون‌ها و مقادیر بحرانی متفاوتی مربوط به کشف مشاهده‌های دورافتاده مورد استفاده قرار گرفته، که برخی از آن‌ها را می‌توان در مورد توزیع نرمال نیز به کار گرفت. آزمون‌هایی که در مورد توزیع نرمال ارائه شده، جداول مقادیر بحرانی برای انواع آزمون‌ها را در حالتی که یک یا هر دو پارامتر توزیع نرمال (میانگین و واریانس) مجهول باشند، شامل می‌شوند (به‌عنوان مثال، تود (۱۹۸۵) را ببینید).

در این نوشتار فقط بررسی شده است که آیا تعداد مشخصی از مشاهده‌ها در یک نمونه‌ی نرمال (مثلاً k تایی آن‌ها) نسبت به بقیه‌ی مشاهده‌ها دورافتاده هستند یا خیر. خواننده‌ی علاقه‌مند به مطالب مربوط به توزیع‌های دیگر می‌تواند به بارنت و لوییس (۱۹۹۴) و تود (۲۰۰۲) مراجعه کند. به‌عنوان یک روش برای بررسی این‌که چه تعداد از مشاهده‌ها در یک نمونه از توزیع نرمال، دورافتاده هستند، می‌توان از آزمون نسبت درست‌نمایی استفاده کرد. برای حالت‌های $k = 1, 2$ توزیع دقیق آماری آزمون نسبت درست‌نمایی مشخص است، اما زمانی که تعداد مشاهده‌های دورافتاده بیش از دو مورد باشد ($k > 2$) هنوز توزیع دقیق این آماره مشخص نیست و این مسئله هنوز بدون جواب باقی مانده است (ژانگ و یو، ۲۰۰۶). هر چند به دست آوردن توزیع دقیق، امری ضروری و لازم به نظر می‌رسد، علی‌رغم تلاش برای یافتن توزیع دقیق، استفاده از تقریب‌های مطلوب نیز روش دیگری است که مناسب به نظر می‌رسد.

در این مقاله، ضمن بیان توزیع تقریبی ارائه‌شده از پیش، برای آماری آزمون نسبت درست‌نمایی (در حالتی که تعداد مشاهده‌های دورافتاده از دو مورد بیشتر باشد) یک تقریب جدید برای توزیع این آماره در حالت کلی مطرح و با تقریب بارنت و لوییس (۱۹۹۴) مقایسه شده است. به‌منظور جامعیت مقایسه‌ی تقریب ارائه‌شده در این مقاله، از دو رویکرد سود جست‌ایم: یکی این‌که مقادیر حاصل از تقریب ارائه‌شده در این مقاله را با مقادیری که از توزیع دقیق برای حالت‌های $k = 1, 2$ حاصل شده‌اند تطابق داده‌ایم، و دیگری این‌که مقادیر حاصل از این تقریب را با تقریب بارنت و لوییس برای موارد $k = 3, 4$ مقایسه کرده‌ایم.

۲ آزمون نسبت درست‌نمایی برای تشخیص k مشاهده‌ی دورافتاده‌ی بالایی (پایینی) در یک نمونه‌ی نرمال

فرض کنید x_1, \dots, x_n نمونه‌ای تصادفی از یک جامعه‌ی نرمال، و $x_{(1)}, \dots, x_{(n)}$ مقادیر مرتب‌شده‌ی آن‌ها باشد. برای تشخیص k مقدار دورافتاده‌ی بالایی در این نمونه، فرض صفر به‌صورت

$H_0: x_1, \dots, x_n \sim N(\mu, \sigma^2)$ و فرض مقابل به صورت زیر است:

$$(۱) \quad H_1: \begin{cases} x_{(1)}, \dots, x_{(n-k)} \sim N(\mu, \sigma^2) \\ x_{(n-k+1)}, \dots, x_{(n)} \sim N(\mu + a, \sigma^2) \end{cases}$$

که در این فرض‌ها پارامترهای μ ، a و σ^2 نامعلوم اند. برای بررسی ناسازگاری k مشاهده‌ی دورافتاده‌ی بالایی با بقیه‌ی مشاهده‌ها، می‌توان آزمون زیر را انجام داد:

$$(۲) \quad \begin{cases} H_0: a = 0 \\ H_1: a > 0 \end{cases}$$

آماره‌ی آزمون نسبت درست‌نمایی برای بررسی k مشاهده‌ی دورافتاده‌ی بالایی ($T_n^{(k)}$) و k مشاهده‌ی دورافتاده‌ی پایینی ($T_{n,(k)}$) به صورت زیر است (بارنت و لویس، ۱۹۹۴):

$$T_n^{(k)} = \frac{x_{(n)} + \dots + x_{(n-k+1)} - k\bar{x}}{s}, \quad T_{n,(k)} = \frac{k\bar{x} - x_{(1)} - \dots - x_{(k)}}{s},$$

که در آن، \bar{x} و s به ترتیب میانگین و انحراف معیار نمونه هستند. پس بنا بر آزمون نسبت درست‌نمایی، هرگاه $T_n^{(k)} \geq C_\alpha$ ، فرض صفر رد می‌شود، که در آن C_α (سطح آزمون است) باید با توجه به توزیع $T_n^{(k)}$ مشخص شود.

۳ توزیع $T_n^{(k)}$ و $T_{n,(k)}$

ابتدا توجه کنید که $(x_{(n)}, \dots, x_{(1)}) \stackrel{d}{=} (-x_{(1)}, \dots, -x_{(n)})$ و بنا بر این $T_n^{(k)} \stackrel{d}{=} T_{n,(k)}$. در نتیجه بدون از دست دادن کلیت مسئله، کافی است فقط در باره‌ی توزیع $T_n^{(k)}$ صحبت کنیم. همچنین به سادگی می‌توان دید که به ازای $k = 1, \dots, n-1$ همواره داریم

$$(۳) \quad T_n^{(k)} \stackrel{d}{=} T_n^{(n-k)}, \quad T_n^{(k)} \stackrel{d}{=} T_{n,(n-k)}, \quad T_{n,(k)} \stackrel{d}{=} T_{n,(n-k)}.$$

توزیع $T_n^{(k)}$ به طور دقیق برای حالتی که فقط یک مشاهده‌ی دورافتاده وجود داشته باشد ($k=1$) با استفاده از روابط بازگشتی به دست آمده است (بارنت و لویس، ۱۹۹۴). با تعمیم روش در حالتی که فقط یک مشاهده‌ی دورافتاده وجود دارد، توزیع $T_n^{(k)}$ با استفاده از یک رابطه‌ی بازگشتی مشابه برای حالتی که فقط دو مشاهده‌ی دورافتاده ($k=2$) در نمونه وجود داشته باشد نیز به دست آمده است (ژانگ و یو، ۲۰۰۶). اما در حالت کلی، هنوز توزیع این آماره در صورت وجود بیش از دو مشاهده‌ی دورافتاده مشخص نیست.

۳/۱ توزیع دقیق $T_n^{(2)}$ و $T_n^{(1)}$

ابتدا توزیع $T_n^{(k)}$ را برای $k = 1, 2$ بیان می‌کنیم. در حالتی که $n = 2$ است، به راحتی می‌توان دید که

$$x_{(2)} = \frac{x_1 + x_2 + |x_1 - x_2|}{2}, \quad s = \frac{|x_1 - x_2|}{\sqrt{2}},$$

و در نتیجه،

$$T_{\frac{1}{\sqrt{2}}}^{(1)} = \frac{x_{(2)} - \bar{x}}{s} = \frac{1}{\sqrt{2}}.$$

بنا بر این، برای $n = 2$ توزیع $T_n^{(1)}$ به صورت زیر است.

$$P\left(T_n^{(1)} \leq t\right) = \begin{cases} 0, & t < \frac{1}{\sqrt{2}} \\ 1, & t \geq \frac{1}{\sqrt{2}} \end{cases}$$

ژانگ و یو (۲۰۰۶) نشان داده‌اند که توزیع $T_n^{(1)}$ برای $n = 3, 4, \dots$ بر اساس رابطه‌ی زیر به دست می‌آید.

$$(۴) \quad P\left(T_n^{(1)} \leq t\right) = n \int_{\frac{1}{\sqrt{n}}}^t P\left(T_{n-1}^{(1)} \leq g_n(x)\right) f_{T_n}(x) dx,$$

که در این رابطه،

$$f_{T_n}(x) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)} \cdot \frac{\sqrt{n}}{n-1} \left(1 - \frac{n}{(n-1)^2} x^2\right)^{\frac{n-2}{2}-1},$$

$$g_n(x) = \frac{nx}{(n-1)\sqrt{\frac{n-1}{n-2} \left\{1 - \frac{nx^2}{(n-1)^2}\right\}}},$$

$$\frac{1}{\sqrt{n}} \leq t \leq \frac{(n-1)}{\sqrt{n}}.$$

به همین ترتیب، رابطه‌ی زیر برای توزیع $T_n^{(2)}$ به دست می‌آید.

$$(۵) \quad P\left(T_n^{(2)} \leq t\right) = P\left(T_n^{(1)} \leq \frac{t}{2}\right) + n \int_{\frac{t}{2}}^{\frac{n-1}{\sqrt{n}}} P\left(T_{n-1}^{(1)} \leq \tilde{g}_n(x)\right) f_{T_n}(x) dx,$$

که در آن،

$$\tilde{g}_n(x) = \frac{t - \frac{n-1}{n-2}x}{(n-1)\sqrt{\frac{n-1}{n-2}\left\{1 - \frac{nx^2}{(n-1)^2}\right\}}},$$

$$\frac{2}{\sqrt{n}} \leq t \leq \sqrt{\frac{2(n-1)(n-2)}{n}}.$$

۳٫۲ تقریب‌هایی برای توزیع $T_n^{(k)}$

محاسبه‌ی توزیع دقیق $T_n^{(k)}$ با استفاده از روابط بازگشتی یاد شده در بخش ۳٫۱ کار ساده‌ای نیست و همان‌طور که پیش‌تر بیان شد، هنوز کاری جدی در این زمینه صورت نگرفته است. در این مقاله دو روش تقریبی برای توزیع $T_n^{(k)}$ مطرح می‌شود؛ یکی از آن‌ها به وسیله‌ی بارتنت و لوییس (۱۹۹۴) و مورد دیگر برای اولین بار در این مقاله اراکه شده است. ابتدا فرض می‌کنیم مجموعه‌ی اندیس‌های $I = \{1, 2, \dots, n\}$ به دو قسمت، یکی $[I]_k = \{i_1, \dots, i_k\}$ با شرط $1 \leq i_1 < i_2 < \dots < i_k \leq n$ و دیگری $(I)_k = I \setminus [I]_k$ افزاز شده است. همچنین نماد $T_{[I]_k}$ را به صورت زیر تعریف می‌کنیم.

$$T_{[I]_k} = \frac{\sum_{i \in [I]_k} X_i - k\bar{X}}{S}.$$

با تعریف این نماد، می‌توان نوشت

$$T_n^{(k)} = \max_{[I]_k} T_{[I]_k}.$$

پس می‌توان گفت که تابع توزیع تجمعی هر یک از $\binom{n}{k}$ متغیر تصادفی $T_{[I]_k}$ ، یک کران بالا برای تابع توزیع تجمعی $T_n^{(k)}$ است.

قضیه‌ی ۱ توزیع متغیر تصادفی $T_n^{(k)}$ به صورت زیر است.

$$f_{T_{[I]_k}}(x) = \frac{\Gamma(\frac{n-1}{r})}{\Gamma(\frac{1}{r})\Gamma(\frac{n-2}{r})} \sqrt{\frac{n}{k(n-k)(n-1)}} \left(1 - \frac{n}{k(n-k)(n-1)}x^2\right)^{\frac{n-2}{r}-1},$$

که در آن،

$$|x| \leq \sqrt{\frac{k(n-k)(n-1)}{n}}.$$

برهان. ابتدا توجه کنید که

$$T_{[I]_k} = \frac{\sqrt{n-1} \frac{k(n-k)}{n} (\bar{X}_k - \bar{X}_{n-k})}{\sqrt{\sum_{i \in [I]_k} (X_i - \bar{X}_k)^2 + \sum_{i \in (I)_k} (X_i - \bar{X}_{n-k})^2 + \frac{k(n-k)}{n} (\bar{X}_k - \bar{X}_{n-k})^2}}.$$

در نتیجه،

$$(۶) \quad \frac{n}{k(n-k)(n-1)} T_{[I]_k}^2 = \left\{ 1 + \frac{\sum_{i \in [I]_k} (X_i - \bar{X}_k)^2 + \sum_{i \in (I)_k} (X_i - \bar{X}_{n-k})^2}{\frac{k(n-k)}{n} (\bar{X}_k - \bar{X}_{n-k})^2} \right\}^{-1},$$

که در روابط بالا $\bar{X}_k = \sum_{i \in [I]_k} X_i / k$ و $\bar{X}_{n-k} = \sum_{i \in (I)_k} X_i / (n-k)$ می‌توان دید که مقدار ماکسیمم طرف راست (۶) برابر با ۱ است و در نتیجه،

$$-\sqrt{\frac{k(n-k)(n-1)}{n}} \leq T_{[I]_k} \leq \sqrt{\frac{k(n-k)(n-1)}{n}},$$

که کران بالا برای $X_i = a$ و $(i \in [I]_k)$ ، $X_i = b \neq a$ و $(i \in (I)_k)$ ، a و b دو عدد حقیقی دلخواه هستند) حاصل می‌شود. همچنین کران پایینی برای $(i \in [I]_k)$ ، $X_i = a$ و $(i \in (I)_k)$ ، $X_i = 2a$ حاصل می‌شود. واضح است که عبارت مخرج در سمت راست (۶) مستقل از صورت است و در ضمن، صورت و مخرج کسر دارای توزیع خی دو، به ترتیب با درجه‌های آزادی $n-2$ و 1 هستند. بنا بر این، نسبت آن‌ها دارای توزیع فیشر با درجه‌های آزادی $n-2$ و 1 است و بنا بر این، کل عبارت دارای توزیع بتا است؛ یعنی،

$$\frac{n}{k(n-k)(n-1)} T_{[I]_k}^2 \sim \text{Beta} \left(\frac{1}{2}, \frac{n-2}{2} \right).$$

نتیجه با یک تبدیل مناسب به دست می‌آید.

۳/۲/۱ تقریب بارنت و لوییس

ابتدا توجه کنید که اگر A_1, A_2, \dots, A_m پیشامدهای دلخواهی باشند، همواره رابطه‌ی زیر برای آن‌ها برقرار است.

$$P \left(\bigcap_{k=1}^m A_k \right) \leq P(A_k), \quad k = 1, 2, \dots, m$$

با استفاده از این نابرابری به راحتی می‌توان دید که

$$P\left(T_n^{(k)} \leq t\right) = P\left(\max_{[I]_k} T_{[I]_k} \leq t\right) = P\left(\bigcap_{[I]_k} \{T_{[I]_k} \leq t\}\right) \leq P\left(T_{[I]_k} \leq t\right)$$

و بنا بر این، یک کران بالا برای توزیع $T_n^{(k)}$ به صورت زیر است.

$$(۷) \quad P\left(T_n^{(k)} \leq t\right) \leq \int_L^t f_{T_{[I]_k}}(x) dx,$$

که در آن،

$$L = -\sqrt{\frac{k(n-k)(n-1)}{n}}, \quad |t| \leq -L$$

همچنین اگر A_1, A_2, \dots, A_m پیشامدهای دلخواهی باشند، بنا بر نابرابری بون فرونی نتیجه می‌شود که

$$P\left(\bigcap_{k=1}^m A_k\right) \geq \sum_{k=1}^m P(A_k) - m + 1$$

و لذا یک کران پایین برای توزیع $T_n^{(k)}$ به صورت زیر به دست می‌آید.

$$(۸) \quad P\left(T_n^{(k)} \leq t\right) \geq m \int_L^t f_{T_{[I]_k}}(x) dx - m + 1, \quad m = \binom{n}{k}$$

این کران می‌تواند به عنوان یک تقریب برای توزیع $T_n^{(k)}$ معرفی شود. این کران، معادل با تقریبی است که برای توزیع $T_n^{(k)}$ در بارنت و لوییس (۱۹۹۴) معرفی شده است. این کران، یک مشکل بزرگ دارد و آن هم این است که یک تابع احتمال نیست و بنا بر این، برای برخی از مقادیر t مناسب نیست. به عنوان مثال، اگر $P(T_{[I]_k} \leq t) \leq (m-1)/m$ باشد، مقدار این کران منفی می‌شود و بنا بر این، استفاده از این تقریب بی معنا است.

۳/۲/۲ تقریب جدید

برای معرفی تقریب جدید، نخست دو مسئله به صورت مجزا بررسی می‌شود. در ابتدا بررسی می‌شود که آیا توزیع مجانبی $T_{[I]_k}$ ها نرمال است و سپس بررسی می‌شود که آیا همبستگی بین $T_{[I]_k}$ ها نیز (هنگامی که اندازه‌ی نمونه زیاد می‌شود) به صفر گرایش دارد. نتیجه‌ی توأم این دو نکته، بررسی فرض استقلال

$T_{[I]_k}$ ‌هاست. برای مورد اول (نرمال بودن توزیع $T_{[I]_k}$ ‌ها) نتیجه‌ی زیر به راحتی به دست می‌آید.

$$\lim_{n \rightarrow \infty} \frac{\Gamma\left(\frac{n-1}{\nu}\right)}{\Gamma\left(\frac{1}{\nu}\right) \Gamma\left(\frac{n-\nu}{\nu}\right)} \sqrt{\frac{n}{k(n-k)(n-1)}} \left(1 - \frac{n}{k(n-k)(n-1)} x^\nu\right)^{\frac{n-\nu}{\nu}-1} \\ = \frac{1}{\sqrt{\nu \pi k}} \exp\left\{-\frac{x^\nu}{\nu k}\right\}.$$

برای توضیح همبستگی بین $T_{[I]_k}$ ‌ها نخست توجه کنید که

$$E\{T_{[I]_k}\} = 0, \quad E\left(\frac{n}{k(n-k)(n-1)} T_{[I]_k}^\nu\right) = \frac{1}{n-1}.$$

در نتیجه،

$$\text{var}(T_{[I]_k}) = E\left(T_{[I]_k}^\nu\right) = \frac{k(n-k)}{n}.$$

همچنین

$$\sum_{[I]_k} T_{[I]_k} = 0, \quad T_{[I]_k} = T_{\{i_1, \dots, i_k\}} = \sum_{j=1}^k T_{i_j}.$$

فرض کنید $k=1$. در این حالت نتیجه می‌شود که

$$\text{var}\left(\sum_{i=1}^n T_i\right) = n-1 + n(n-1)\text{cov}(T_{\{i\}}, T_{\{j\}}) = 0.$$

در نتیجه،

$$\rho(T_{\{i\}}, T_{\{j\}}) = \frac{-1}{n-1}, \quad i \neq j$$

با توجه به این رابطه، با افزایش n ، همبستگی بین $T_{[I]_k}$ ‌ها به صفر نزدیک می‌شود. اما همه‌ی همبستگی‌های بین $T_{[I]_k}$ ‌ها برای k ‌های بزرگ‌تر از ۱ لزوماً منفی نیستند. برای حالت $k=2$ نتیجه می‌شود که

$$\text{cov}(T_{\{i_1, i_2\}}, T_{\{j_1, j_2\}}) = \text{cov}(T_{\{i_1\}} + T_{\{i_2\}}, T_{\{j_1\}} + T_{\{j_2\}}) = \text{cov}(T_{\{i_1\}}, T_{\{j_1\}}) \\ + \text{cov}(T_{\{i_1\}}, T_{\{j_2\}}) + \text{cov}(T_{\{i_2\}}, T_{\{j_1\}}) + \text{cov}(T_{\{i_2\}}, T_{\{j_2\}})$$

$$= \begin{cases} -\frac{4}{n}, & i_1 \neq i_2 \neq j_1 \neq j_2 \\ \frac{4}{n}, & \begin{cases} i_1 = j_1 \neq i_2 \neq j_2 \\ i_1 = j_2 \neq i_2 \neq j_1 \\ i_2 = j_1 \neq i_1 \neq j_2 \\ i_2 = j_2 \neq i_1 \neq j_1 \end{cases} \end{cases}$$

بنا بر این، ضریب همبستگی به صورت زیر است.

$$\rho(T_{\{i_1, i_2\}}, T_{\{j_1, j_2\}}) = \begin{cases} -\frac{2}{n-1}, & i_1 \neq i_2 \neq j_1 \neq j_2 \\ \frac{n-4}{2n-4}, & \begin{cases} i_1 = j_1 \neq i_2 \neq j_2 \\ i_1 = j_2 \neq i_2 \neq j_1 \\ i_2 = j_1 \neq i_1 \neq j_2 \\ i_2 = j_2 \neq i_1 \neq j_1 \end{cases} \end{cases}$$

برای حالت کلی، فرض کنید z نشان دهنده تعداد اندیس‌های برابر در زوج $(T_{[I]_k}, T_{[J]_k})$ یا مانند نمادگذاری قبلی، در زوج $(T_{\{i_1, \dots, i_k\}}, T_{\{j_1, \dots, j_k\}})$ باشد. در این صورت، ضریب همبستگی بین $T_{\{i_1, \dots, i_k\}}$ و $T_{\{j_1, \dots, j_k\}}$ از رابطه زیر به دست می‌آید.

$$(9) \quad \rho(T_{\{i_1, \dots, i_k\}}, T_{\{j_1, \dots, j_k\}}) = \frac{zn - k^2}{kn - k}, \quad z = 0, \dots, k-1$$

در نتیجه،

$$(10) \quad \lim_{n \rightarrow \infty} \rho(T_{\{i_1, \dots, i_k\}}, T_{\{j_1, \dots, j_k\}}) = \frac{z}{k}, \quad z = 0, \dots, k-1$$

از طرفی، توزیع فراوانی z در جدول زیر داده شده است.

z	۰	۱	۲	...	$k-1$	کل
f_i	$\binom{n}{k} \binom{k}{1} \binom{n-k}{k}$	$\binom{n}{k} \binom{k}{1} \binom{n-k}{k-1}$	$\binom{n}{k} \binom{k}{2} \binom{n-k}{k-2}$...	$\binom{n}{k} \binom{k}{k-1} \binom{n-k}{1}$	$\binom{n}{k}^2 - \binom{n}{k}$

با توجه به جدول توزیع فراوانی z ، نسبت مقادیر z برابر با صفر، به کل تعداد زوج‌ها، بر اساس رابطه‌ی

زیر به سمت ۱ میل می‌کند.

$$\frac{\binom{n}{k} \binom{n-k}{k}}{\binom{n}{k}^2 - \binom{n}{k}} = \frac{\binom{n-k}{k}}{\binom{n}{k} - 1} \geq \frac{\binom{n-k}{k}}{\binom{n}{k}} = \frac{(n-k)!(n-k)!}{n!(n-2k)!} \rightarrow 1$$

در نتیجه $۱ \rightarrow P(Z = ۰)$ و این بدان معنا است که با افزایش n ، متغیرهای تصادفی $\{T_{[I]_k}\}$ به سمت ناهمبستگی و نرمال بودن میل می‌کنند و لذا فرض استقلال متغیرهای تصادفی $\{T_{[I]_k}\}$ تا اندازه‌ای قابل قبول است.

با توجه به آنچه بیان شد، ما تقریب زیر را برای توزیع $T_n^{(k)}$ در حالت کلی پیشنهاد می‌کنیم.

$$(۱۱) \quad \begin{aligned} P\left(T_n^{(k)} \leq t\right) &= P\left(\max_{[I]_k} T_{[I]_k} \leq t\right) = P\left(\bigcap_{[I]_k} \{T_{[I]_k} \leq t\}\right) \\ &\approx \left\{ \int_L^t f_{T_{[I]_k}}(x) dx \right\}^{\binom{n}{k}}, \quad k = ۱, ۲, \dots, n-۱ \end{aligned}$$

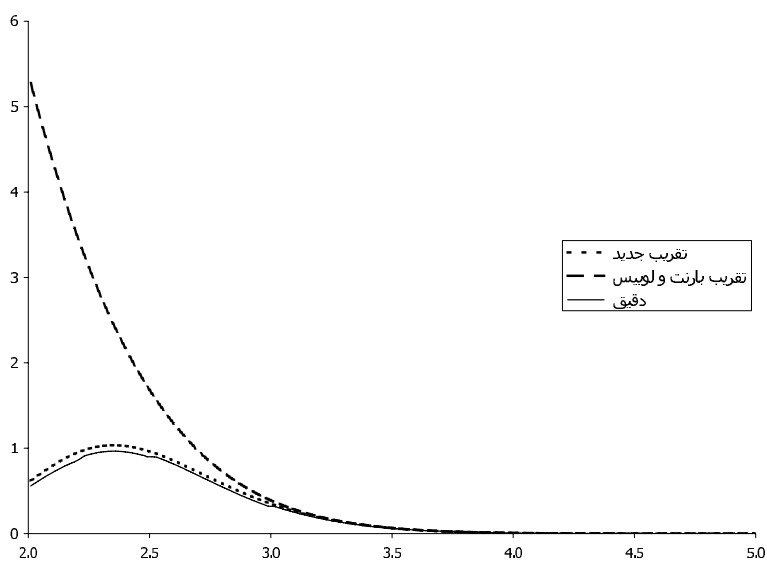
۳٫۳ ویژگی‌های تقریب جدید و بررسی دقت آن

به منظور بررسی دقت تقریب پیشنهاد شده، ذکر چند نکته لازم به نظر می‌رسد. یکی این‌که تقریب ارائه شده خاصیت هم‌توزیعی‌ای را که برای حالت دقیق، طبق رابطه‌ی (۳) وجود دارد، دارا است؛ زیرا به راحتی می‌توان دید که $T_{[I]_k} \stackrel{d}{=} T_{[I]_{n-k}}$ و در نتیجه،

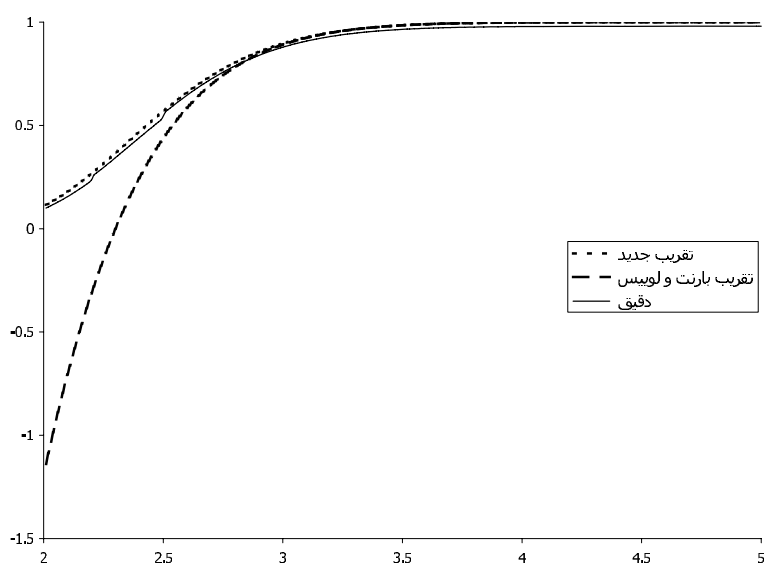
$$(۱۲) \quad \begin{aligned} P\left(T_n^{(n-k)} \leq t\right) &= P\left(\max_{[I]_{n-k}} T_{[I]_{n-k}} \leq t\right) = P\left(\bigcap_{[I]_{n-k}} \{T_{[I]_{n-k}} \leq t\}\right) \\ &\approx \left\{ \int_L^t f_{T_{[I]_{n-k}}}(x) dx \right\}^{\binom{n}{n-k}} \\ &= \left\{ \int_L^t f_{T_{[I]_k}}(x) dx \right\}^{\binom{n}{k}}, \quad k = ۱, \dots, n-۱. \end{aligned}$$

از (۱۱) و (۱۲) نتیجه می‌شود که توزیع‌های تقریبی معرفی شده برای $T_n^{(k)}$ و $T_n^{(n-k)}$ یکسان‌اند. همچنین تقریب پیشنهادی در این مقاله، عیب تقریب مطرح شده در بارنت و لویس (۱۹۹۴) را ندارد؛ یعنی، یک تابع احتمال است. به عنوان مثال، با استفاده از نرم‌افزار Maple نمودارهایی برای مقایسه‌ی تابع توزیع تجمعی و تابع چگالی احتمال رسم شده است (شکل‌های ۱ و ۲).

در شکل ۱، بخشی از تابع چگالی احتمال $T_{\cdot}^{(۱)}$ برای حالت دقیق و با استفاده از تقریب بارنت و لویس و تقریب جدید رسم شده است. در این شکل، محور افقی مقدار t را نشان می‌دهد و محور عمودی مقدار تابع چگالی احتمال $f_{T_{\cdot}^{(۱)}}(t)$ را. مطابق با شکل ۱ دیده می‌شود که تقریب جدید، بسیار نزدیک‌تر به تابع چگالی دقیق است تا تقریب



شکل ۱. تابع‌های چگالی دقیق و تقریبی برای $T_{1..}^{(1)}$



شکل ۲. تابع‌های توزیع تجمعی دقیق و تقریبی برای $T_{1..}^{(1)}$

بارنت و لوییس. همچنین دقت تقریب‌ها با نزدیک شدن به دنباله‌ی توزیع تقریباً برابر می‌شود. اما واضح است که در حالت کلی، دقت تقریب بارنت و لوییس در برآورد تابع چگالی احتمال، نسبت به تقریب جدید بسیار کم‌تر است.

در شکل ۲، بخشی از تابع توزیع تجمعی $T_{\cdot\cdot}^{(1)}$ برای حالت دقیق و با استفاده از تقریب بارنت و لوییس و تقریب جدید رسم شده است. در این شکل، محور افقی مقدار t را نشان می‌دهد و محور عمودی مقدار تابع توزیع تجمعی $P(T_{\cdot\cdot}^{(1)} \leq t)$ را.

در شکل ۲ دیده می‌شود که تقریب جدید برای t های کوچک، بسیار نزدیک‌تر به تابع توزیع تجمعی دقیق است تا تقریب بارنت و لوییس. به تدریج با بزرگ شدن مقدار t ، تفاوت سه منحنی کم می‌شود. اما واضح است که در حالت کلی، دقت تقریب بارنت و لوییس در برآورد تابع توزیع تجمعی، نسبت به تقریب جدید بسیار کم‌تر است.

نکته‌ی دوم، دقت یا اختلاف مقادیر بحرانی به دست آمده از تقریب بارنت و لوییس و تقریب جدید با مقادیری است که از روش دقیق به دست آمده‌اند. با توجه به وجود توزیع دقیق $T_n^{(k)}$ برای $k = 1, 2$ ، مقادیرهای بحرانی به صورت دقیق محاسبه شده‌اند (ژانگ و یو، ۲۰۰۶) و همین مقادیرها با استفاده از روش‌های تقریبی محاسبه شده‌اند؛ لذا می‌توان آن‌ها را با هم مقایسه و دقت تقریب بارنت و لوییس و تقریب جدید را بررسی کرد. در جدول‌های ۱ و ۲ برخی از مقادیرهای بحرانی مربوط به توزیع $T_n^{(k)}$ (در حالت‌های $k = 1, 2$) که حاصل روش‌های تقریبی‌اند، همراه با مقادیرهای دقیق محاسبه‌شده‌ی آن‌ها – با استفاده از روابط (۱۱) و (۱۲) – ارائه شده است. در توصیف نتایج مندرج در این دو جدول، چهار نکته‌ی زیر جالب توجه است.

آ) تقریب جدید از لحاظ دقت، بهتر از تقریب بارنت و لوییس است؛ هر چند در برخی موارد، اختلاف آن‌ها ناچیز است. البته باید متذکر شد چنان‌چه $y = \int_L^t f_{T_{[1]k}}(x) dx$ باشد، به راحتی می‌توان دید که

$$0 \leq y \leq 1 \quad \Rightarrow \quad y^m \leq y, \quad y^m \geq my - m + 1$$

و بنا بر این، دقت تقریب جدید، بیش‌تر از دقت تقریب بارنت و لوییس است.

ب) دقت تقریب جدید برای حالت $k = 1$ بیش‌تر است و قطعاً بدان دلیل است که همبستگی‌ها در این حالت به صفر نزدیک‌ترند و سریع‌تر همگرا به صفر می‌شوند.

جدول ۱. مقدارهای بحرانی برای توزیع $T_n^{(k)}$ در سطح $\alpha = 0.1$

n						
۱۰۰	۵۰	۳۰	۲۰	۱۰	۵	
						$k = 1$
۳,۶۰۰	۳,۳۳۷	۳,۱۰۳	۲,۸۸۴	۲,۴۱۰	۱,۷۴۹	دقیق
۳,۶۵۰	۳,۳۴۰	۳,۱۰۶	۲,۸۸۶	۲,۴۱۳	۱,۷۵۱	تقریب بارت و لوییس
۳,۶۰۰	۳,۳۳۵	۳,۱۰۲	۲,۸۸۳	۲,۴۱۰	۱,۷۴۹	تقریب جدید
						$k = 2$
۶,۱۱۸	۵,۴۹۷	۴,۹۴۶	۴,۴۳۷	۳,۴۰۲	۲,۱۶۰	دقیق
۶,۱۹۳	۵,۵۵۲	۴,۹۸۰	۴,۴۶۵	۳,۴۰۶	۲,۱۶۴	تقریب بارت و لوییس
۶,۱۳۶	۵,۵۱۶	۴,۹۵۱	۴,۴۳۵	۳,۴۰۲	۲,۱۶۰	تقریب جدید

جدول ۲. مقدارهای بحرانی برای توزیع $T_n^{(k)}$ در سطح $\alpha = 0.5$

n						
۱۰۰	۵۰	۳۰	۲۰	۱۰	۵	
						$k = 1$
۳,۲۰۷	۲,۹۵۶	۲,۷۴۵	۲,۵۵۷	۲,۱۷۶	۱,۶۷۱	دقیق
۳,۲۲۵	۲,۹۶۷	۲,۷۴۹	۲,۵۵۹	۲,۱۷۹	۱,۶۷۲	تقریب بارت و لوییس
۳,۲۰۷	۲,۹۵۵	۲,۷۴۳	۲,۵۵۳	۲,۱۷۳	۱,۶۷۰	تقریب جدید
						$k = 2$
۵,۶۳۸	۵,۰۵۸	۴,۵۶۱	۴,۱۱۰	۳,۱۹۷	۲,۱۰۱	دقیق
۵,۷۴۵	۵,۱۲۳	۴,۶۰۰	۴,۱۲۳	۳,۱۹۸	۲,۱۰۳	تقریب بارت و لوییس
۵,۶۵۷	۵,۰۹۴	۴,۵۸۴	۴,۱۱۳	۳,۱۹۳	۲,۱۰۰	تقریب جدید

پ) دقت هر دو تقریب برای n های کوچک، بسیار زیاد است و برای n های بزرگ تر، تقریب جدید دقیق تر است.

ت) با افزایش α دقت تقریب ها کم می شود تا جایی که همان طور که قبلاً مطرح شد، برای برخی از مقدارهای α تقریب بارت و لوییس غیر قابل استفاده است.

در جدول های ۳ و ۴ مقدارهای بحرانی شبیه سازی شده (با استفاده از نمونه های به اندازه ی $10,000$)، حاصل از برنامه ی S-Plus، که ضمیمه ی مقاله است) برای $k = 3, 4$ ، همراه با مقدارهای محاسبه شده از

جدول ۳. مقدارهای بحرانی برای توزیع $T_n^{(k)}$ در سطح $\alpha = 0.01$

n					
۱۰۰	۵۰	۳۰	۲۰	۱۰	
					$k = 3$
۸,۳۸۸	۷,۳۲۹	۶,۴۳۱	۵,۶۱۲	۳,۹۹۷	شبیه‌سازی
۸,۴۹۳	۷,۳۹۹	۶,۴۶۷	۵,۶۳۰	۴,۰۰۴	تقریب بارتنت و لوییس
۸,۴۷۴	۷,۳۸۸	۶,۴۵۱	۵,۶۱۴	۳,۹۹۸	تقریب جدید
					$k = 4$
۱۰,۳۰۹	۸,۹۳۵	۷,۶۶۰	۶,۵۳۰	۴,۳۲۳	شبیه‌سازی
۱۰,۶۲۱	۹,۰۵۲	۷,۷۱۷	۶,۵۴۴	۴,۳۳۱	تقریب بارتنت و لوییس
۱۰,۵۹۹	۹,۰۳۵	۷,۷۰۰	۶,۵۲۹	۴,۳۲۳	تقریب جدید

جدول ۴. مقدارهای بحرانی برای توزیع $T_n^{(k)}$ در سطح $\alpha = 0.05$

n					
۱۰۰	۵۰	۳۰	۲۰	۱۰	
					$k = 3$
۷,۸۵۵	۶,۸۷۱	۶,۰۵۱	۵,۳۱۱	۳,۸۱۳	شبیه‌سازی
۸,۰۵۶	۶,۹۹۹	۶,۱۱۱	۵,۳۲۱	۳,۸۱۸	تقریب بارتنت و لوییس
۸,۰۴۴	۶,۹۹۲	۶,۱۰۰	۵,۳۱۴	۳,۸۱۴	تقریب جدید
					$k = 4$
۹,۷۷۲	۸,۴۰۸	۷,۲۳۵	۶,۲۴۹	۴,۱۵۵	شبیه‌سازی
۱۰,۱۸۲	۸,۶۶۱	۷,۳۷۹	۶,۲۶۱	۴,۱۶۱	تقریب بارتنت و لوییس
۱۰,۱۷۲	۸,۶۵۱	۷,۳۷۰	۶,۲۵۳	۴,۱۵۵	تقریب جدید

روش‌های تقریبی ارائه شده است. در مقایسه‌ی مقدارهای بحرانی در جدول‌های ۳ و ۴ به نظر می‌رسد که با افزایش n ، مقدارهای بحرانی تا حدودی نسبت به مقدارهای شبیه‌سازی شده بزرگ‌تر باشند، در حالی که تفاوت آن‌ها برای نمونه‌های دارای اندازه‌ی کم‌تر از 30 بسیار ناچیز است. البته باید توجه داشت که در یک تعداد تکرار ثابت، برای n های کوچک‌تر، روش شبیه‌سازی دقیق‌تر عمل می‌کند. به هر ترتیب به نظر می‌رسد که تقریب بارتنت و لوییس و تقریب جدید، برای n های کوچک نیز بسیار مطلوب‌اند و نسبت به روش دقیق، نیاز به محاسباتی به مراتب کم‌تر دارند. علاوه بر این، مانند حالت $k = 1, 2$ دقت تقریب جدید، بیش‌تر از دقت تقریب بارتنت و لوییس است و با افزایش n ، این تفاوت بیش‌تر می‌شود.

۴ بحث و نتیجه‌گیری

در حدود یک قرن است که تشخیص مشاهده‌های دورافتاده مورد توجه آمارشناسان و دیگر دانشمندان است. محققان در طی سالیان متمادی روش‌های متعددی را برای تشخیص مشاهده‌های دورافتاده معرفی کرده‌اند. به دلیل اهمیت توزیع نرمال در مباحث آمار کاربردی، این مسئله برای توزیع نرمال نیز مطرح شده است. در این مقاله برای تشخیص مشاهده‌های دورافتاده در یک نمونه‌ی استخراج شده از توزیع نرمال، آزمون نسبت درست‌نمایی مورد استفاده قرار گرفت. توزیع دقیق آماری آزمون نسبت درست‌نمایی به دست آمده، هنوز در نوشتگان آماری به جز برای یک و دو مشاهده‌ی دورافتاده نامشخص است. در این مقاله برای توزیع صفر آماری آزمون نسبت درست‌نمایی در حالت کلی، یک فرمول تقریبی جدید (رابطه‌ی (۱۱) را ببینید) معرفی و ویژگی‌های آن بررسی شد. از مقایسه‌ی مقدارهای حاصل شده از تقریب پیشنهادی، تقریب بارتنت و لوییس، و توزیع دقیق برای یک و دو مشاهده‌ی دورافتاده دیده شد (جدول‌های ۱ و ۲ را ببینید) که دقت تقریب پیشنهاد شده در این مقاله بسیار زیاد است. همچنین در مقایسه با مقدارهای شبیه‌سازی شده (که در حال حاضر، مورد استفاده‌ی عملی است) در حالتی که ۳ یا ۴ مشاهده‌ی دورافتاده در نمونه وجود داشته باشد، دیده شد (جدول‌های ۳ و ۴ را ببینید) که دقت تقریب بارتنت و لوییس و تقریب جدید، زیاد است و البته برای اندازه‌ی نمونه‌ی بزرگ‌تر، مقدارهای محاسبه شده از روش‌های تقریبی، اندکی بزرگ‌تر می‌شوند. همچنین دیده شد (جدول‌های ۱ تا ۴ را ببینید) که دقت تقریب‌ها با حرکت به سمت دنباله‌ی راست توزیع $T_n^{(k)}$ (یعنی برای α های کوچک‌تر) رو به افزایش است. اما با وجود این که به نظر می‌رسد دقت تقریب بارتنت و لوییس و تقریب جدید، چندان متفاوت نیست، دیده شد که در برخی موارد، تقریب بارتنت و لوییس منجر به نتایج نادرست می‌شود. در واقع، همان‌طور که پیش‌تر اشاره شد، تقریب بارتنت و لوییس، تابع احتمال نیست و نمی‌توان از آن برای برآورد تابع چگالی احتمال استفاده کرد؛ در حالی که با استفاده از تقریب جدید، این امر نیز میسر است. نکته‌ی دیگری که در باره‌ی این تقریب‌ها می‌توان مطرح کرد، ویژگی محاسباتی آن‌ها است که برای حالت‌های $k = 1, 2$ به مراتب منعطف‌تر و کم‌تر از روش دقیق است.

مرجع‌ها

- Aggarwal, C.C.; Yu, P.S. (2001). Outlier detection for high dimensional data. In Proc. ACM SIGMOD.
- Barnett, V.; Lewis, T. (1994). *Outliers in Statistical Data*, 3rd ed. Wiley, Chichester.
- Chambers, J.; Cleveland, W.; Kleiner, B.; Tukey, P. (1983). *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA.

- David, H.A. (1981). *Order Statistics*, 2nd ed. Wiley, New York.
- Ferguson, T.S. (1961). On the rejection of outliers. Proc. Berkeley Symp. Math. Statist. and Prob. 1 Univ. of Cla. Press, Berkeley, pp. 253-285.
- Grubbs, F. (1969). Procedures for detecting outlying observation in samples. *Technometrics* **11**, 1-12.
- Hampel, F.R.; Ronchetti, E.; Rousseeuw, P.J.; Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.
- Hawkins, D.M. (1980). *The Identification of Outliers*. Chapman, London.
- Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.
- Johnson, B.A.; Hunt, H.H. (1979). Performance characteristics for certain tests to detect outliers. Proceeding of statistical computing section, Annual meeting of the american statistical association, Washington, B.C.
- Moore, D.S.; McCabe, G.P. (1999). *Introduction to the Practice of Statistics*, 3rd ed. New York.
- Rousseeuw, P.J.; Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.
- Thode, H.C. Jr. (1985). Power of absolute moment tests against symmetric non-normal alternatives. Ph.D. Dissertation, University Microfilms, Ann Arbor, Michigan.
- Thode, H.C. Jr. (2002). *Testing for Normality*. Marcel Dekker, New York.
- Thode, H.C. Jr.; Smith, L.A.; Finch, S.J. (1983). Power of test of normality for detecting scale contaminated normal samples. *Commun. Stat. Simul. Comput.* **12**, 675-695.
- Zhang, J.; Yu, k. (2006). The null distribution of the likelihood ratio test for one or two outliers in a normal sample. *Test* **15**, 141-150.

حسین حسینی
 بانک مرکزی جمهوری اسلامی ایران،
 خیابان پاسداران، ساختمان شماره ۱ (۲۰۷)،
 تهران، ایران.
 گروه آمار، دانشکده ریاضی،
 دانشگاه کاردیف،
 کاردیف، انگلستان.
 پیام‌نگار: hassanih@cf.ac.uk

رحیم محمودوند
 گروه آمار،
 دانشگاه پیام نور مرکز تویسرکان،
 تویسرکان، ایران.
 پیام‌نگار: r_mahmodvand@yahoo.com

پیوست

برنامه‌ی زیر به راحتی می‌تواند برای شبیه‌سازی مقدارهای برخی از چندک‌های توزیع $T_n^{(k)}$ مورد استفاده قرار گیرد. در این برنامه، $rept$ تعداد تکرارها برای شبیه‌سازی، $start$ کم‌ترین مقدار ممکن n ، و end مقدار پایانی مورد نظر برای n است.

```
quanoutnk<-function(rept,k,start,end){
z<-array(dim=c(end-start+1,6))
header<-c("alpha=0.20","alpha=0.10","alpha=0.05","alpha=0.01","alpha=0.005",
"alpha=0.001")
lmargin<-c(paste("n="start:end))
dimnames(z)<-list(lmargin,header)
for(n in start : end){
y<-array(dim=rept)
for(i in 1 : rept){
x<-sort(c(rnorm(n)))
y[i]<-(sum(x[(n-k+1) : n])-k*mean(x))/sqrt(var(x))
z[n-start+1,1 :6]<-quantile(y,c(0.8,0.9,0.95,0.99,0.995,0.999))}}
z}
```